The Astrophysical Journal, 862:101 (18pp), 2018 August 1 https://doi.org/10.3847/1538-4357/aacdaa

A Machine-learning Method for Identifying Multiwavelength Counterparts of Submillimeter Galaxies: Training and Testing Using AS2UDS and ALESS

Fang Xia An[1,2,3], S. M. Stach[3], Ian Smail[3], A. M. Swinbank[3], O. Almaini[4], C. Simpson[5], W. Hartley[4], D. T. Maltby[4], R. J. Ivison[6,7], V. Arumugam[6,7], J. L. Wardlow[3], E. A. Cooke[3], B. Gullberg[3], A. P. Thomson[8], Chian-Chou Chen[6], J. M. Simpson[9], J. E. Geach[10], D. Scott[11], J. S. Dunlop[7], D. Farrah[12,13], P. van der Werf[14], A. W. Blain[15], C. Conselice[4], M. Michałowski[16], S. C. Chapman[17], and K. E. K. Coppin[10]

[1] Purple Mountain Observatory, China Academy of Sciences, 8 Yuanhua Road, Nanjing 210034, People's Republic of China; fangxiaan@pmo.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[3] Centre for Extragalactic Astronomy, Department of Physics, Durham University, Durham, DH1 3LE, UK
[4] University of Nottingham, School of Physics and Astronomy, Nottingham, NG7 2RD, UK
[5] Gemini Observatory, Northern Operations Center, 670 N. A'ohuku Place, Hilo, HI 96720, USA
[6] European Southern Observatory, Karl Schwarzschild Strasse 2, Garching, Germany
[7] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[8] The University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[9] Academia Sinica Institute of Astronomy and Astrophysics, No. 1, Section 4, Roosevelt Rd., Taipei 10617, Taiwan
[10] Centre for Astrophysics Research, School of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield AL10 9AB, UK
[11] Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, BC V6T 1Z1, Canada
[12] University of Hawaii, 2505 Correa Road, Honolulu, HI 96822, USA
[13] Virginia Polytechnic Institute and State University, Department of Physics, MC 0435, 910 Drillfield Drive, Blacksburg, VA 24061, USA
[14] Leiden Observatory, Leiden University, P.O. Box 9513, NL-2300RA Leiden, the Netherlands
[15] Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK
[16] Astronomical Observatory Institute, Faculty of Physics, Adam Mickiewicz University, ul. Sloneczna 36, 60-286 Poznań, Poland
[17] Department of Physics and Atmospheric Science, Dalhousie University, Halifax, NS B3H 3J5, Canada
Received 2018 February 9; revised 2018 May 18; accepted 2018 June 17; published 2018 July 27

Abstract

We describe the application of supervised machine-learning algorithms to identify the likely multiwavelength counterparts to submillimeter sources detected in panoramic, single-dish submillimeter surveys. As a training set, we employ a sample of 695 ($S_{870\mu m} \gtrsim 1$ mJy) submillimeter galaxies (SMGs) with precise identifications from the ALMA follow-up of the SCUBA-2 Cosmology Legacy Survey's UKIDSS-UDS field (AS2UDS). We show that radio emission, near-/mid-infrared colors, photometric redshift, and absolute *H*-band magnitude are effective predictors that can distinguish SMGs from submillimeter-faint field galaxies. Our combined radio + machine-learning method is able to successfully recover ∼85% of ALMA-identified SMGs that are detected in at least three bands from the ultraviolet to radio. We confirm the robustness of our method by dividing our training set into independent subsets and using these for training and testing, respectively, as well as applying our method to an independent sample of ∼100 ALMA-identified SMGs from the ALMA/LABOCA ECDF-South Survey (ALESS). To further test our methodology, we stack the 870 μm ALMA maps at the positions of those *K*-band galaxies that are classified as SMG counterparts by the machine learning but do not have a $>4.3\sigma$ ALMA detection. The median peak flux density of these galaxies is $S_{870\mu m} = (0.61 \pm 0.03)$ mJy, demonstrating that our method can recover faint and/or diffuse SMGs even when they are below the detection threshold of our ALMA observations. In future, we will apply this method to samples drawn from panoramic single-dish submillimeter surveys that currently lack interferometric follow-up observations to address science questions that can only be tackled with large statistical samples of SMGs.

Key words: cosmology: observations – galaxies: evolution – galaxies: formation – galaxies: high-redshift – galaxies: starburst – submillimeter: galaxies

Supporting material: machine-readable table

1. Introduction

The bulk of the population of submillimeter-luminous galaxies (SMGs) are massive, dust-enshrouded systems that are forming stars at rates of $\gtrsim 10^2 - 10^3\,M_\odot\,\mathrm{yr}^{-1}$. At these star formation rates (SFRs), these systems would, in principle, be able to form the stellar mass of massive galaxies ($M_* \gtrsim 10^{11}\,M_\odot$) within just ∼100 Myr (e.g., Chapman et al. 2005; Bothwell et al. 2013; Casey et al. 2014). Although such strongly star-forming galaxies are rare in the local universe, the space density of bright SMGs (i.e., $S_{850\mu m} > 1$ mJy, corresponding to a far-infrared luminosity, $L_{\rm IR} \gtrsim 10^{12}\,L_\odot$) increases rapidly with look-back time and appears to peak at $z \sim 2$–3 (e.g., Barger et al. 1999; Chapman et al. 2005; Smolčić et al. 2012; Yun et al. 2012; Simpson et al. 2014). Due to their potentially rapid formation, SMGs have been proposed to be the progenitors of spheroidal galaxies in the local universe (e.g., Lilly et al. 1999; Swinbank et al. 2006; Simpson et al. 2014, 2017). They are also thought to be linked to quasi-stellar object (QSO) activity due the similarity of their redshift distribution to that of luminous QSOs (e.g., Coppin et al. 2008), as well as being linked to compact red galaxies seen at $z \sim 1$–2 (e.g., Cimatti et al. 2008; Whitaker et al. 2012; Toft et al. 2014). These characteristics mean that SMGs may be an important stage in the formation and evolution of massive galaxies and hence are a key element to constrain models of galaxy formation and evolution.

Submillimeter/millimeter galaxy selection benefits from the strong negative *k*-correction in these wavebands (Blain & Longair 1993), which enables us to detect sources above a constant flux limit and hence with near-constant SFRs out to high redshift ($z \sim 6$). In the past two decades, numerous wide-field submillimeter surveys have been undertaken on the James Clerk Maxwell Telescope (JCMT), IRAM 30 m, APEX, and ASTE equipped with the SCUBA/SCUBA-2, MAMBO, LABOCA, and AZTEC cameras, respectively (e.g., Smail et al. 1997; Barger et al. 1998; Hughes et al. 1998; Scott et al. 2002, 2012; Coppin et al. 2006; Weiß et al. 2009; Ikarashi et al. 2011; Geach et al. 2017; Wang et al. 2017; and see Casey et al. 2014 for a review). The main challenge for follow-up studies of the sources selected from these surveys is the coarse angular resolution of the single-dish maps, with the full width at half maximum (FWHM) typically around $\sim$8″–10″ at 450 μm (but only for relatively small surveys; Geach et al. 2013; Wang et al. 2017) and $\sim$15″–20″ in the wide-field surveys undertaken at 850–1100 μm (Weiß et al. 2009; Geach et al. 2017), which results in uncertain identifications of the counterparts at other observed frequencies.

Traditionally, the likely counterparts for single-dish submillimeter sources were identified by using indirect tracers of the far-infrared/submillimeter emission, such as the radio, 24 μm, or mid-infrared properties (e.g., Ivison et al. 1998; Smail et al. 2002; Pope et al. 2006; Ivison et al. 2007; Barger et al. 2012; Michałowski et al. 2012; Cowie et al. 2017). These properties roughly track the far-infrared luminosity of galaxies, and they have two additional advantages: that observations in these bands are typically at significantly higher angular resolution than the submillimeter, and that the surface densities of sources in these wavebands are relatively low, so that the rate of chance associations is also low. Unfortunately, the negative *k*-correction experienced in the submillimeter band arises from the steeply rising Rayleigh–Jeans part of the spectral energy distribution (SED), the absence of which in these other wavebands means that even the deepest radio continuum or mid-infrared maps will miss the highest-redshift SMGs. Nevertheless, $\sim$50% of submillimeter sources can be located via a radio or mid-infrared identified counterpart (e.g., Ivison et al. 2002, 2007, 2010; Hodge et al. 2013). To improve on this situation and so construct more complete samples of SMGs, it is necessary to combine a broader range of multiwavelength properties to isolate potential SMGs from the less active galaxies that are found within the error circles of single-dish submillimeter sources (e.g., Chapin et al. 2011; Alberts et al. 2013; Chen et al. 2016). One additional complication of these statistical identifications is the fact that recent studies using interferometric observations in the submillimeter/millimeter suggest that $\gtrsim$20% of single-dish-detected submillimeter sources actually correspond to blends of multiple SMGs (e.g., Wang et al. 2011; Karim et al. 2013; Simpson et al. 2015a, 2015b; Stach et al. 2018a).

Recently, interferometric observations undertaken at submillimeter/millimeter wavelengths with the Atacama Large Millimeter/submillimeter Array (ALMA) are helping to improve our understanding of SMGs. With angular resolution better than 1″ and thus subarcsecond positional precision, we are starting to obtain a more complete understanding of the multiwavelength characteristics of SMGs (e.g., Hodge et al. 2013; Swinbank et al. 2014, 2015; Thomson et al. 2014; Aravena et al. 2016; Walter et al. 2016; Danielson et al. 2017; Dunlop et al. 2017; Simpson et al. 2017; Wardlow et al. 2017). However, for single-dish submillimeter surveys of fields in the northern sky, it is not possible to perform ALMA follow-up, and so we must rely instead on the use of the Submillimeter Array (SMA) or IRAM's Northern Extended Millimetre Array (NOEMA) to obtain interferometric identifications (e.g., Smolčić et al. 2012; Hill et al. 2018). Moreover, for very large samples of submillimeter sources, it may be challenging to obtain complete identifications even with ALMA.

The rapid growth of data from panoramic single-dish submillimeter surveys (Geach et al. 2017; Wang et al. 2017; J. M. Simpson et al. 2018, in preparation) requires the adoption of fast, automatic techniques for identifying the likely counterparts to single-dish-detected submillimeter sources. Automatized classification using machine-learning algorithms has recently gained popularity in astronomy and has been applied to a number of problems, including star/galaxy/quasar classification (Bloom et al. 2012; Solarz et al. 2012; Małek et al. 2013; Kurcz et al. 2016) and the identification of different types of supernovae (du Buisson et al. 2015; Lochner et al. 2016).

In this work, we test two machine-learning algorithms, support vector machine (SVM) and extreme gradient boosting (XGBoost), to identify probable SMG counterparts from optical/near-infrared-selected galaxies.

The SVMs are a class of supervised-learning algorithms based on the structural risk minimization principle developed by Vapnik (1995). The main idea behind support vector classification (SVC) is to determine decision planes between sets of objects with different class labels and then calculate a decision boundary by maximizing the margin between the closest points of the classes. Each single object is then classified based on its relative position in a multidimensional parameter space.

The second machine-learning algorithm we test is XGBoost (Chen & Guestrin 2016), which is a modified version of gradient boosting (Friedman 2001) used for supervised-learning problems. The basic model of XGBoost is a tree ensemble, which is a set of classification and regression trees. In this model, each input feature of an object will be divided into different "leaves," and each "leaf" will be assigned a score. This score will be used as a quality on a tree structure. A greedy algorithm that starts from a single leaf and iteratively adds branches to the tree is used to construct the structure of a tree. In this gradient-boosting tree model, one of the basic functions is to search for an optimal split at each node. To make this decision, XGBoost calculates the structure score of all possible splits and finds the best solution among them. In practice, multiple trees will be used together to be trained on the properties of objects in the training set, and the final prediction will be made by summing the scores in the corresponding leaves of each individual tree in the tree ensemble model (Chen & Guestrin 2016).

Generally, there are four steps to performing a supervised machine-leaning classification: (1) construct a training set, (2) identify the optimal features that can best separate different classes, (3) train the machine-learning models to build a classifier, and (4) apply to the test sample to classify the unknown objects.

In this work, we exploit the multiwavelength counterparts of ~700 ALMA-detected SMGs identified by S. M. Stach et al. (2018b, in preparation) in their ALMA follow-up of the SCUBA-2 Cosmology Legacy Survey (S2CLS; Geach et al. 2017) observations of the UKIRT Infrared Deep Sky Survey (UKIDSS; Lawrence et al. 2007) Ultra Deep Survey (UDS) field (O. Almaini et al. 2018, in preparation). We begin by identifying counterparts to ALMA SMGs by matching them to a deep *K*-band-selected photometric catalog of the UKIDSS-UDS field (O. Almaini et al. 2018, in preparation; W. Hartley et al. 2018, in preparation). We then compare the multi-wavelength properties of the SMGs and a sample of non-SMG field galaxies (which lie within the footprint of our ALMA observations but are individually undetected in these sensitive submillimeter maps) and identify those properties that can best separate these two populations. We train the machine-learning classifiers based on these selected properties to construct a method to identify probable SMG counterparts for single-dish-detected submillimeter sources that are not yet or cannot be observed with ALMA. By utilizing our method, we can construct larger and more robust samples of counterparts to SMGs that can be used to answer the science questions related to the evolutionary cycle of SMGs and their connections with other populations.

Given the proven success of radio observations in locating counterparts to a subset of the SMG population, we adopt a two-pronged approach, where we combine a simple probability cut to select likely radio counterparts, followed by a machine-learning method applied to multiwavelength data to increase the completeness of the resulting SMG sample. We choose to apply these two selections separately, rather than combining the radio fluxes into the machine-learning analysis, primarily because of the requirements in terms of multiwavelength detections needed for the SVM machine-learning analysis. As we show, applying the radio and SVM machine-learning classifications independently maximizes the completeness of the final SMG sample.

The plan of this paper is as follows. We introduce the observations of the training set we use in the S2CLS UDS field and an independent test sample from the Extended *Chandra* Deep Field South (ECDFS) in Section 2. Our methodology is described in Section 3. We present and discuss our results in Section 4. The main conclusions of this work are given in Section 5. Throughout this paper, we adopt a cosmology with $[\Omega_\Lambda, \Omega_M, h_{70}] = [0.7, 0.3, 1.0]$. The AB magnitude system (Oke 1974) is used unless otherwise stated.

2. Observational Training Set and Test Sample

2.1. ALMA-identified Sample of SMGs

To construct our training set and the test sample, we employ two wide-field, single-dish submillimeter surveys that have been uniformly followed up using ALMA in the same submillimeter band as the original surveys (to remove ambiguity in the identification of counterparts). These then provide us with a sample of SMGs with a wide range of properties and submillimeter fluxes, and, equally importantly, they yield samples of field galaxies that fall within the ALMA survey footprint but are undetected in those maps and hence can be used as a control sample of submillimeter-faint galaxies to try to distinguish the unique characteristics of SMGs.

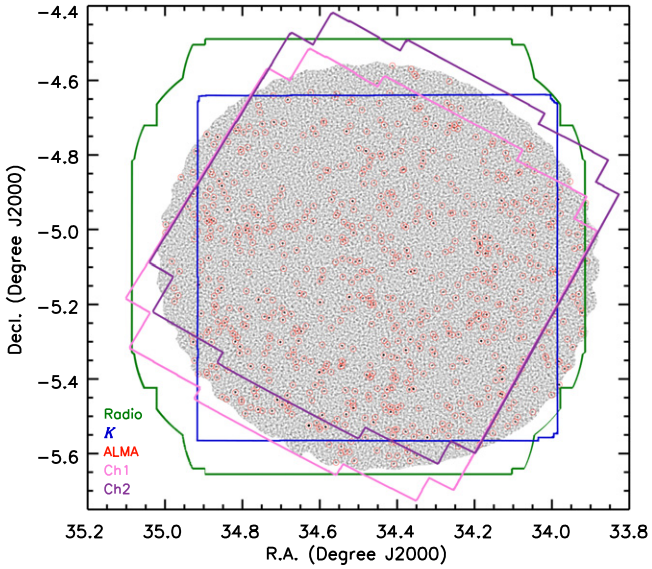

Figure 1. Map showing the distribution of our ALMA survey compared to the coverage of the *K*-band, *Spitzer*, and VLA observations of the UDS field and overlaid on the SCUBA-2 map. We circle the positions of our 716 ALMA pointings. All but the most western two ALMA pointings are covered by the radio map. In addition, 643/716 (~90%) of the ALMA pointings fall within the deepest UKIDSS near-infrared coverage. High-quality photometric redshifts are available for those sources within the overlap region of the UKIDSS and *Spitzer* IRAC 3.6 μm (Ch 1) and 4.5 μm (Ch 2) imaging. Of the ALMA pointings in this region, 607/716 (~85%) are suitable for using as a training set for our machine-learning method. We therefore limit our machine-learning analysis to this region.

2.1.1. Single-dish Sample

The UKIDSS-UDS field (R.A./decl.: 02h, −05°; Figure 1) was mapped with the SCUBA-2 bolometer camera (Holland et al. 2013) on the JCMT at 850 μm as part of the S2CLS. We provide a brief overview here; the full details of observations, data reduction, and catalog are described in Geach et al. (2017). The coverage of 0.96 deg^2 in UDS is relatively uniform, with instrumental noise varying by only ~5% across the field (Figure 3 in Geach et al. 2017). The final matched-filtered map has a noise of $\leqslant$1.3 mJy beam^{-1} rms over 0.96 deg^2 and 0.82 mJy beam^{-1} in the deepest part. The empirical point-spread function (PSF) has an FWHM of 14.″8. Geach et al. (2017) identified a total of 716 submillimeter sources above a 4σ limit with a false-detection rate of ~2% (Figure 13 in Geach et al. 2017).

We also employ a second single-dish survey sample in our analysis as an additional test of our method. This sample comprises the 126 submillimeter sources with a single-to-noise ratio (S/N) > 3.7 from the LABOCA ECDFS submillimeter Survey (LESS; Weiß et al. 2009) taken with the Atacama Pathfinder Experiment (APEX) telescope. This 870 μm map covers 0.25 deg^2 with a 19.″2 FWHM and a 1σ depth of $S_{870\mu m} = 1.2$ mJy. The properties of this sample are thus similar to those of the S2CLS UDS sample but in a completely independent field with different multiwavelength coverage and photometric selection. We refer the reader to Weiß et al. (2009) for the details of these observations.

2.1.2. ALMA Follow-up

Band 7 (870 μm) observations have been obtained with ALMA of all 716 submillimeter sources from the S2CLS UDS map, which are described in full in S. M. Stach et al. (2018b, in preparation). Observations of 30 of the brightest ($S_{850\mu m} \geqslant 8$ mJy) single-dish sources were undertaken in Cycle 1 as part of a pilot project, 2012.1.00090.S (Simpson et al. 2015a, 2015b, 2017), while observations of the bulk of the sample were obtained through the Cycle 3 project 2015.1.01528.S and the Cycle 4 project 2016.1.00434.S. The Cycle 1 pilot observations relied on an early interim map, and in the 30 ALMA maps, 52 SMGs were detected at $\geqslant 4\sigma$ significance (Simpson et al. 2015a, 2015b). However, in the final SCUBA-2 maps, three of these 30 sources fall below our sample selection criteria, leaving 27 of them in our final sample of single-dish-detected submillimeter sources. In Cycles 3 and 4, we observed the remaining 686 sources with $S_{850\mu m} \geqslant 3.5$ mJy from the final S2CLS map (S. M. Stach et al. 2018b, in preparation). These observations achieve typical 1σ depths of $\sigma_{870\mu m} \sim 0.25$ mJy with synthesized beams of 0.″15–0.″3. The ALMA maps are tapered to $\sim$0.″5 resolution before sources are identified. Across all 716 single-dish submillimeter sources, we detect 695 SMGs above $>4.3\sigma$ (corresponding to a false-detection rate of 2%). We refer to our complete 870 μm ALMA survey of 716 SCUBA-2 sources in the UDS field as the "AS2UDS" survey. We note that the ALMA primary beam of our observation is 17.″4, which encompasses the area of the SCUBA-2 beam. Full details of the observation, data reduction, source detection, and cataloging are presented in S. M. Stach et al. (2018b, in preparation).

Among the 716 ALMA maps, 108 do not contain any ALMA-identified SMGs at $>4.3\sigma$. We label these as "blank-ALMA" maps. In the remaining 608 ALMA maps, we detected 695 SMGs with fluxes from $S_{850\mu m} = 0.89$ to 30 mJy. In the following, these maps are described as "maps with ALMA ID."

The goal of this study is to develop a method to reliably and robustly identify counterparts to single-dish-detected submillimeter sources in wide-field surveys by utilizing the multiwavelength properties of the sample of ALMA-identified SMGs. Therefore, we include the multiwavelength galaxies lying within the 108 "blank-ALMA" maps in our analysis to guarantee the completeness of our parent single-dish sample.

In our analysis, we will use independent subsets of the AS2UDS SMG sample to test the reliability of our method. We also include an additional sample for this purpose: the ALMA follow-up of the LESS survey. The ALESS survey obtained ALMA 870 μm observations in Cycle 0 of 122 of the 126 LESS sources (Hodge et al. 2013). These early ALMA observations have a typical synthesized beam of $\sim$1.″6 and 1σ depths of $\sim$0.4 mJy but with a wider range of data quality than the later AS2UDS survey. For this reason, in this work, we only use the 88 "good-quality" ALMA maps from Hodge et al. (2013) to construct our test sample. Again, these include 19 "blank-ALMA" maps, which lack detected SMGs. These 88 maps yield a sample of 96 ALMA-detected SMGs with multiwavelength coverage from Simpson et al. (2014), which we will employ in our analysis. We note that the properties of this test sample differ from those of the AS2UDS sample, as it is based on an IRAC-selected photometric catalog, as opposed to the K band for AS2UDS, and the photometric redshifts are derived using different codes in the two fields. This comparison is intended to illustrate the results that would be obtained if a training set from one field were simply applied directly to a sample selected from a second survey, with different selection and photometric coverage.

2.2. Multiwavelength Observations

We next describe the multiwavelength observations of the UDS and ECDFS fields that are used to determine the properties of our SMG samples. We will focus on the radio and redder optical and near-infrared bands, as the dusty, star-forming SMGs are expected to be typically brighter in these wavebands than the bulk of the field population (e.g., Wardlow et al. 2011; Michałowski et al. 2012; Hodge et al. 2013; Simpson et al. 2014).

2.2.1. VLA Observations

Since radio synchrotron emission arises from supernova remnants, it provides a powerful tracer of obscured star formation. As such, radio emission has been traditionally used to identify counterparts to SMGs (e.g., Ivison et al. 1998, 2002).

In this work, we exploit the VLA observations of the UDS at 1.4 GHz (21 cm), which were carried out by the UDS20 survey (V. Arumugam et al. 2018, in preparation). These VLA observations cover an area of 1.3 deg^2 centered on the UDS field. The typical rms noise across the full VLA map is 10 μJy, and it is 7 μJy beam^{-1} at its deepest point in the center. In total, 6861 radio sources are detected above 4σ. The details of the observations, data reduction, and catalog will be discussed in V. Arumugam et al. (2018, in preparation). In total, 714/716 ALMA pointings fall within the VLA map (Figure 1).

2.2.2. Optical/Near-infrared Observations in the UDS

Deep near-infrared imaging data are crucial for investigating the properties of SMGs because of their high redshifts and dusty nature. The UKIDSS-UDS represents one of the deepest near-infrared imaging surveys over a wide area, covering 0.8 deg^2. As shown in Figure 1, $\sim$90% (643/716) of our ALMA pointings are covered by UKIDSS.

The near-infrared image we exploit in our analysis is taken from UDS Data Release 11 (DR11; O. Almaini et al. 2018, in preparation), which represents the final UDS release over the whole field. Details of observations, data reduction, and catalog extraction will be presented in the forthcoming UDS data paper (O. Almaini et al. 2018, in preparation). Briefly, the DR11 reaches 3σ median depths of $J = 26.2$, $H = 25.7$, and $K = 25.9$ mag, which are measured in a 2″ diameter aperture. In total, 296,007 sources were detected from the K-band image using SExtractor (Bertin & Arnouts 1996), with the photometry in the J and H bands obtained in SExtractor dual-image mode.

The Y-band data are from the Visible and Infrared Survey Telescope for Astronomy (VISTA) Deep Extragalactic Observations (VIDEO) survey with 3σ depths of $Y = 25.3$ mag (Jarvis et al. 2013). The optical B-, V-, R_c-, i'-, and z'-band observations of UDS were carried out using the Suprime-Cam on the Subaru telescope (Furusawa et al. 2008) with 3σ depths of $B = 28.4$, $V = 27.8$, $R_c = 27.7$, $i' = 27.7$, and $z' = 26.6$ mag in 2″ diameter apertures. The field was also observed by the Megacam on the Canada–France–Hawaii Telescope (CFHT) in the u' band to a 3σ limiting depth of $u' = 27.3$ mag, again in a 2″ diameter aperture.

The mid-infrared observations of the UDS were taken with IRAC and at 24 μm with MIPS by the *Spitzer* Legacy Program (SpUDS; PI: J. Dunlop). The 5σ depths of the IRAC 3.6 and 4.5 μm observations are [3.6] = 24.2 and [4.5] = 24.0 mag.

In total, 12-band data (*UBVRIzYJHK*[3.6][4.5]) are utilized to derive photometric redshifts for the 296,007 *K*-band-detected sources. Details of the photometric-matched catalog and color measurement will be described in W. Hartley et al. (2018, in preparation). Hartley et al. used Easy and Accurate Redshifts from Yale (EAZY; Brammer et al. 2008) to estimate the photometric redshift for the *K*-band-detected sample. To obtain unbiased and high-quality photometric redshifts, they only considered those sources within the joint IRAC (SpUDS) and UKIDSS coverage and excluded those sources that have contaminated photometry (i.e., due to halos from bright stars or other artifacts). In total, ∼85% (607/716) of the ALMA pointings fall in the region for which reliable photometric redshifts are available. Photometric redshifts were derived in the manner described by Simpson et al. (2013; see also Hartley et al. 2013; Mortlock et al. 2013). Hartley et al. compared the estimated photometric redshifts of ∼6500 sources with available spectroscopic redshifts in the DR11 and found that the accuracy of the photometric redshifts is $|z_{\rm spec} - z_{\rm phot}|/(1 + z_{\rm spec}) = 0.019 \pm 0.001$.

2.2.3. Multiwavelength Observations in the ECDFS

The radio, optical, and near-infrared observations of our independent test sample in the ECDFS are presented in Simpson et al. (2014). The VLA 1.4 GHz data used in Simpson et al. (2014) and this work are from Miller et al. (2008). We use the radio catalog from Miller et al. (2008) to identify radio counterparts to IRAC-based galaxies in the ECDFS. Biggs et al. (2011) rereduced the VLA 1.4 GHz imaging data in the ECDFS and created a deep radio catalog containing sources down to an S/N of 3 for searching radio counterparts to single-dish-detected SMGs. We also use this deep radio catalog in our analysis to calculate the completeness of radio identification in the ECDFS. The depth and quality of the multiwavelength coverage of the ECDFS is broadly comparable to that available for the UDS in terms of number and depth of the photometric bands. For detailed information on the depth and coverage of the optical and near-infrared data in the ECDFS, the reader is referred to Table 2 of Simpson et al. (2014).

2.3. Matching SMGs to Multiwavelength Data

As the first step in our analysis, we match the ALMA-identified SMGs to the multiwavelength data from their respective fields and determine the properties of ALMA SMGs based on their multiwavelength counterparts.

2.3.1. Matching to Radio Counterparts in the UDS

Since radio identification has been proven to be an efficient tool to search for counterparts of bright SMGs (e.g., Ivison et al. 2002; Chapman et al. 2005; Hodge et al. 2013), we first match our SMGs to the radio source catalogs. As shown in Figure 1, 714 of 716 ALMA maps in the UDS field are covered by the available VLA observations. There are 404 radio sources (Figure 2) that fall inside the 17.″4 diameter FWHM of the primary-beam coverage of the 714 ALMA maps. To identify probabilistic radio counterparts to the low-resolution, SCUBA-2-detected submillimeter sources, we include all 404 $\geqslant 4\sigma$ radio sources within the ALMA maps in our analyses.

Before matching ALMA SMGs to the radio sources, we first check the cumulative number of matches to obtain an appropriate matching radius between ALMA SMGs and radio sources. A radius of 1.″6 is chosen because the cumulative number of matches becomes flat beyond this radius. Within this matching radius, the false-match rate is ∼1%. From the 695 AS2UDS SMGs, 693 are covered by the VLA radio observations. Among these, 268 ALMA SMGs match to 259 radio sources within 1.″6 (Figure 2), with nine radio sources having two ALMA counterparts. In total, 39% (268/695) of the AS2UDS SMGs have a radio counterpart brighter than the 4σ limit of the VLA catalog.

We then assess the robustness of our $\geqslant 4\sigma$ radio catalog. As we showed above, there are 404 radio sources in the area covered by our ALMA maps. Of these, 259 radio sources are counterparts to ALMA SMGs, along with 42 radio sources that lack both *K*-band and ALMA counterparts (and hence may be spurious). However, using the IRAC coverage of the field, we find that 17 of the 42 have 3.5 and 4.6 μm detections, indicating that about half of these are real radio sources but lack the *K* and ALMA detections. This suggests that the spurious source fraction in our radio catalog is less than 25/404, or $\lesssim$6%. Raising the significance cut on the radio catalog to $\geqslant 5\sigma$ reduces the number of *K*/IRAC/ALMA blank radio sources to 10 (from 310 radio sources, or an upper limit on the spurious fraction of $\lesssim$3%) but also removes 40 radio counterparts to ALMA SMGs and thus reduces the completeness of our identifications. For this reason, we have chosen to retain the $\geqslant 4\sigma$ flux limit on the radio catalog.

To start with, for the SCUBA-2-detected submillimeter sources, we first consider all $\geqslant 4\sigma$ radio sources within the ALMA primary beam as potential counterparts. Then we calculate the corrected-Poissonian probability, *p*-value (Downes et al. 1986; Dunlop et al. 1989), for all 404 radio sources falling in our ALMA maps by using

$$\begin{aligned} E &= P_{\rm c} \quad P^* \geqslant P_{\rm c} \\ E &= P^*\{1 + \ln(P_{\rm c}/P^*)\} \quad P^* \leqslant P_{\rm c}, \end{aligned} \tag{1}$$

where $P_{\rm c}$ is the critical Poisson probability level given by $P_c = \pi r_{\rm s}^2 N_{\rm T}$, in which $N_{\rm T}$ is the surface density of the radio sources and $r_{\rm s}$ is the search radius (in this work, it is the radius of the ALMA primary beam). Then, given P^* for a radio source, we can derive the probability that it is a counterpart of single-dish-detected submillimeter sources by $p = \{1 - \exp(-{\rm E})\}$.

As shown in Figure 2, the fraction of counterparts of ALMA SMGs among the radio sources dramatically decreases when $p > 0.065$. Hence, we adopt $p \leqslant 0.065$ as our limit for the probabilistic association of radio sources to single-dish submillimeter sources, while we consider those radio sources with $p = 0.065$–0.10 as "possible" identifications. Looking at all 404 radio sources falling in our ALMA maps, 41 of these have $p > 0.065$–0.10 and are thus only classified as "possible" counterparts (Figure 2). Of these "possible" counterparts, the vast majority (36/41) do not match to an ALMA-identified SMG. As a result, the five radio sources from these 41 that do match to ALMA SMGs within 1.″6 are also removed by utilizing the *p*-value cut. We also show the spatial offset of SCUBA-2 source positions and radio sources in Figure 2. We see that those radio sources with $p > 0.065$ have spatial offsets

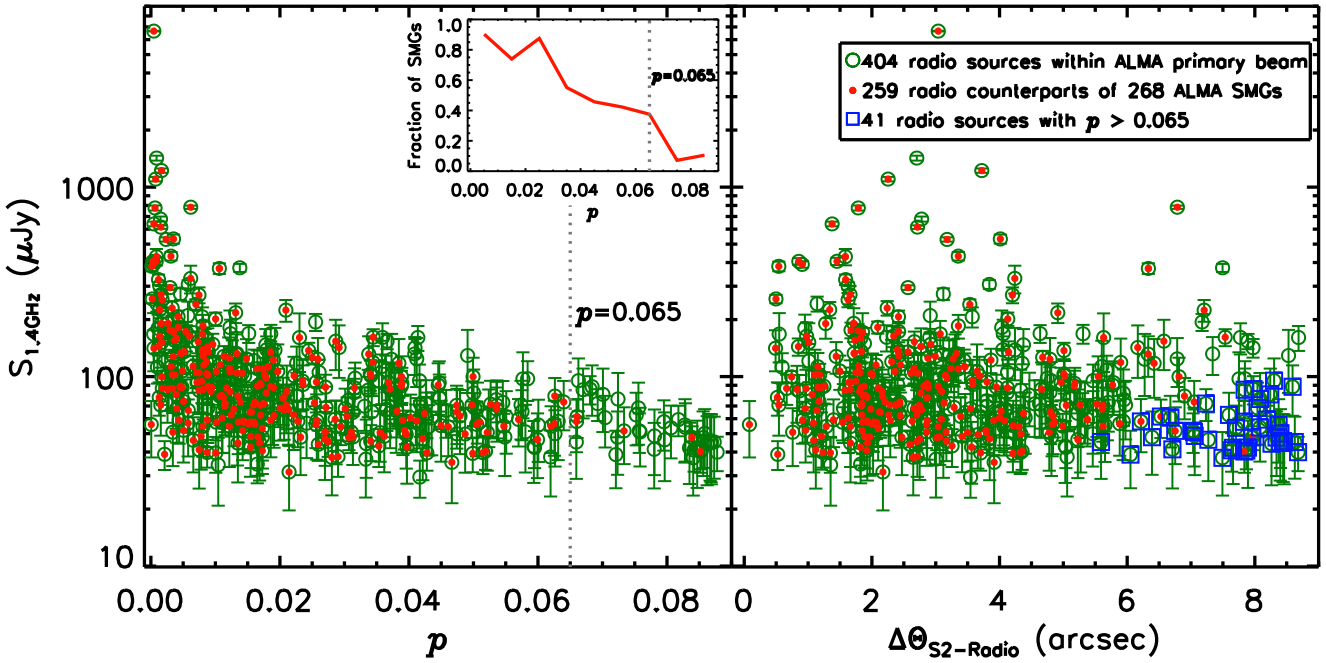

Figure 2. Radio flux densities for all radio sources within the primary beams of the AS2UDS ALMA maps as a function of the corrected-Poissonian probability, p-value (left), and the offset of these radio sources from the SCUBA-2 single-dish source position (right). In total, there are 404 radio sources within the ALMA maps in the UDS (open circles). Among these, 259 radio sources are matched to 268/695 ALMA SMGs within a radius of 1.″6 (filled circles), including nine ALMA SMGs that have double radio counterparts. Hence, ∼63% of radio sources within the ALMA maps correspond to counterparts of ALMA SMGs. We utilize the corrected-Poissonian probability, p-value, to estimate the likelihood of a radio source being the counterpart of a single-dish-detected submillimeter source. We show the fraction of counterparts of ALMA SMGs from all 404 radio sources within the ALMA maps as a function of p-value in the inset plot of the left panel. The number of counterparts of SMGs dramatically decreases when $p > 0.065$. Therefore, we choose $p \leqslant 0.065$ as a cut of "robust" radio identifications in this work. There are 41 radio sources that have $p > 0.065$ (blue squares). The majority of these are not associated with SMGs, so we adopt $p \leqslant 0.065$ as our limit for identifying radio counterparts to SMGs. Using this p-value, the precision of radio identification for identifying counterparts of SCUBA-2-detected SMGs is ∼70%.

larger than 5.″5 from the nominal SCUBA-2 positions. However, if we simply adopt this smaller match radius to search for radio counterparts to SCUBA-2 sources, we will remove ∼20 of the radio counterparts to actual ALMA SMGs. Therefore, in this work, we prefer to consider all radio sources within the ALMA primary beam but apply a $p \leqslant 0.065$ cut to identify those that are likely counterparts to the SCUBA-2-detected submillimeter sources. As a result, the precision of radio identification of counterparts to single-dish-detected sources increases from 64% (259/404) to 70% (254/363) by utilizing this p-value cut. Precision is defined as the ratio between the correctly identified SMGs and the total number of predicted SMGs by radio identification/machine-learning classification.

To identify those multiwavelength properties that differentiate the SMGs from the wider field population, we define radio sources that do not match to an ALMA-detected SMG within 2.″6 (this is conservatively chosen to be larger than our 1.″6 matching radius) as "non-SMG" radio sources. Including the 53 radio sources within the "blank-ALMA" maps, in total, there are 137 non-SMG radio sources falling within our ALMA maps. Although, as we show later, on average, the radio sources within the "blank-ALMA" maps have faint submillimeter emission, we put them into the sample of non-SMGs for simplicity before we perform the stacking analysis. We will discuss the properties of radio sources that are counterparts of SMGs and non-SMGs in Section 4.

2.3.2. Matching to Near-infrared/Optical Counterparts in the UDS

To develop a method to differentiate SMGs and non-SMGs using multiwavelength data, we adopt the UDS DR11 photometric-matched near-infrared/optical catalog (W. Hartley et al. 2018, in preparation) to identify counterparts and measure the near-infrared/optical colors of SMGs.

As we described above, only those sources within the overlapped region of UKIDSS and IRAC have sufficient photometric coverage and estimated photometric redshifts, as well as absolute magnitudes, which we will use in our machine-learning method. Hence, we limit our identification of counterparts to the ALMA SMGs in this region. In total, 607/716 ALMA maps fall in this region, and 583/695 ALMA SMGs are detected within these maps with $K \leqslant 25.9$ mag.

To select a suitable matching radius between K-band galaxies and ALMA SMGs, we test radii between 0.″5 and 1.″0 in steps of 0.″1 and match the K-band galaxies with the ALMA SMGs. At each step, we randomly offset the K-band galaxies in right ascension or declination by 10″–20″ to estimate the false-match fraction as a function of matching radius. At a match radius of 0.″6, 514 K-band galaxies from the UKIDSS DR11 photometric catalog match to ALMA SMGs with a false-match fraction of ∼3.5% (∼18 false matches). A match radius of 0.″5 reduces the false-match fraction to 2% (∼10 false matches) but also reduces the total number of matches by 20. A larger match radius increases the matched sources, but the new matches are dominated by false matches. Therefore, we adopt a match radius of 0.″6.

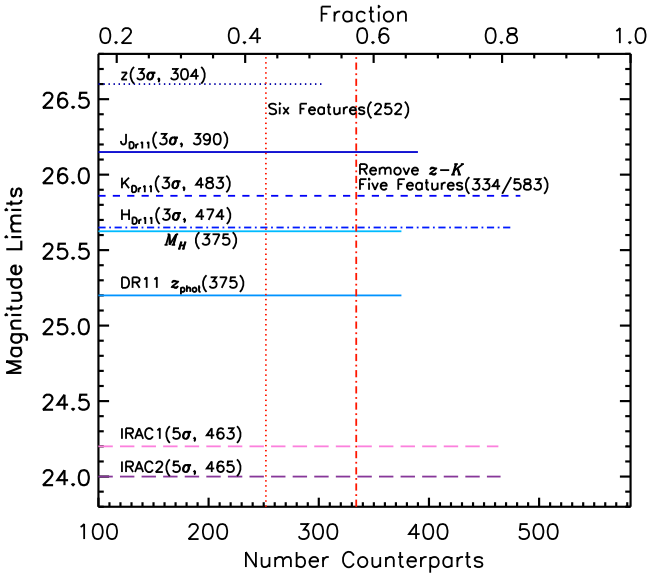

Figure 3. Number of multiwavelength counterparts to ALMA-detected SMGs within the overlap regions of UKIDSS and IRAC coverage in the UDS field. As shown in Figure 1, ∼85% of our ALMA maps are covered by UKIDSS and IRAC observations, and 583/695 ALMA-detected SMGs lie in the combined footprint. The horizontal lines indicate the 3σ (or 5σ) limit of the corresponding photometric band that is used as part of the multiwavelength selection when identifying the counterparts to SMGs. We can see that ∼83% of the ALMA-identified SMGs have a *K*-band counterpart, but the number of detected counterparts dramatically decreases at bluer wavelengths. We also show the number of ALMA-identified SMGs that have a photometric redshift estimate and absolute rest-frame *H*-band magnitude. The vertical lines show the fraction of SMGs that have six features (dotted line; $(z-K)$, $(J-K)$, $(K-[3.6])$, [3.6]–[4.5], $z_{\rm phot}$, and M_H) or five features (dot-dashed line; removing $(z-K)$) that will be used in our machine-learning method.

In the overlap region of UKIDSS and IRAC, there are 483 *K*-band galaxies that match to ALMA SMGs within our adopted 0.″6 matching radius. We show the number and fraction of multiwavelength counterparts of ALMA-detected SMGs in Figure 3. We find that ∼83% (483/583) of the ALMA SMGs have *K*-band counterparts, but the number of counterparts dramatically decreases at bluer wavelengths due to their dusty nature (and their likely high redshifts). For the optical and near-infrared data, we use the 3σ limits to identify the counterparts as shown in Figure 3. Because of the relatively low resolution of the IRAC data, a more conservative 5σ cut is adopted for identifying counterparts and measuring colors in these bands. Figure 3 also presents the number of SMGs that have photometric redshifts, which are estimated based on DR11 photometric catalog, and hence have absolute *H*-band magnitudes available to be used in the following analyses.

2.3.3. Radio and Optical/Near-infrared Counterparts in the ECDFS

The details of the identification of radio and optical/near-infrared counterparts to the ALESS SMGs in the ECDFS field are presented in Hodge et al. (2013) and Simpson et al. (2014), respectively. Out of the 96 ALMA SMGs, 45 have radio counterparts (Hodge et al. 2013). Simpson et al. (2014) measured aperture photometry in 19 wavebands for the 96 ALMA SMGs. Among these, 77 are securely detected and have sufficient photometry to derive a photometric redshift and estimate the rest-frame *H*-band absolute magnitudes.

For the single-dish-detected submillimeter sources, we first use the IRAC-based photometric catalog of sources in the ECDFS from Simpson et al. (2014) to match 88 LESS submillimeter sources (Weiß et al. 2009) for which there are good-quality ALMA maps from Hodge et al. (2013). We include in this the 19 submillimeter sources for which the corresponding ALMA map detected no SMGs (the "blank-ALMA" maps). In total, there are 323 IRAC-detected galaxies located within the 88 ALMA primary beams. We will use these galaxies to test our methodology in the following analysis.

3. Method: Radio + Machine-learning Identifications

To apply supervised machine-learning classification, we require a list of observed properties for a training sample made up of submillimeter-detected and submillimeter-undetected galaxies. Therefore, we first need to select those features of SMGs that best separate them from field galaxies ("non-SMGs"). Given the power of radio identification to locate the counterparts, we adopt a two-pronged approach, where we combine a likelihood test to select probable radio counterparts with a machine-learning method to increase the completeness of the resulting SMG sample. As we will show, we apply these two tests separately, in part because of the requirements in terms of multiwavelength detections needed for the machine-learning analysis and in part because of differences in the coverage of the field in the radio, optical, and near-infrared imaging data sets.

For the machine-learning analysis, we note that previous work has shown that SMGs are in general at high redshift, are relatively bright in the rest-frame near-infrared, and have red colors in the optical and near-infrared wavebands (e.g., Smail et al. 2002; Chapman et al. 2005; Hainline et al. 2009; Michałowski et al. 2012; Wang et al. 2012; Alberts et al. 2013; Simpson et al. 2014; Chen et al. 2016). To compare the properties of the SMGs to the field, we use as our ("non-SMG") control sample those *K*-band-detected sources that are located within the primary beams of our ALMA maps but are >1.″6 away from an ALMA-identified SMG. In total, there are 4658 non-SMG *K*-band galaxies within the ALMA primary beam area (a total area of 47.3 arcmin^2). Among them, 799 lie within the 108 "blank-ALMA" maps.

3.1. "Blank-ALMA" Maps

As we described in Section 2.1.2, we include the 108 "blank-ALMA" maps in our analysis to ensure our tests accurately reflect the success rate of identifying counterparts to "typical" single-dish submillimeter sources. However, due to the ambiguity about the submillimeter emission from those galaxies lying in the "blank-ALMA" maps, we first investigate the average far-infrared emission of these "blank-ALMA" maps before we include them in the sample of "non-SMG" galaxies used to identify the properties that can cleanly differentiate SMGs and non-SMGs and to construct the training set for machine learning.

We note that the false-positive rate (FPR) for the SCUBA-2 catalog (weighted by the number of sources at a given S/N) is ∼2% at $>4\sigma$ (Geach et al. 2017), meaning that we expect around ∼15 of our SCUBA-2 sources to be spurious, with these sources contributing to the 108 "blank-ALMA" maps. To test this, we stack the *Herschel*/SPIRE maps at the position of all 108 "blank-ALMA" maps. We detected significant emission with flux densities of 16.4 ± 0.6, 16.0 ± 0.8, and 10.4 ± 1.0 mJy at 250, 350, and 500 μm, respectively. Adopting the typical 850/500 μm color for SMGs from Swinbank et al. (2014), this corresponds to a typical 850 μm

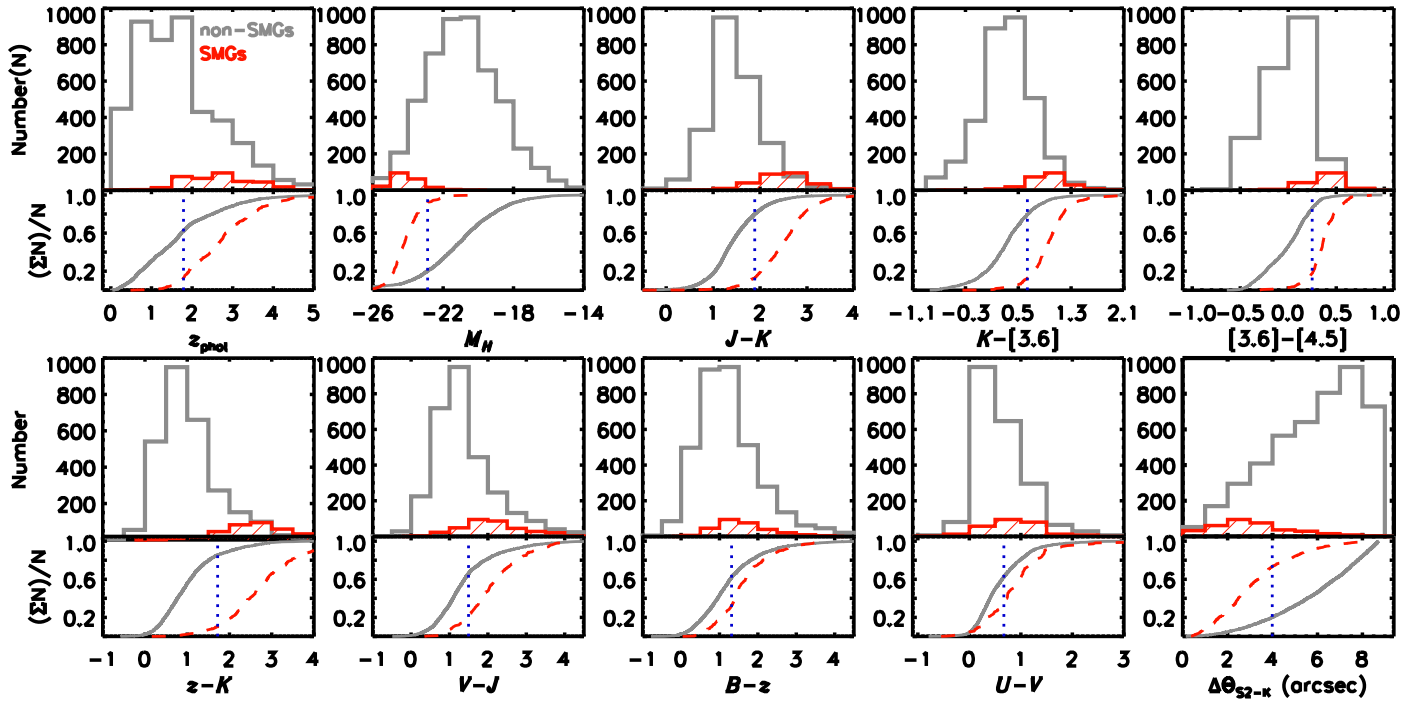

Figure 4. Histograms of different observed properties of SMG vs. non-SMG field galaxies. Non-SMGs are defined as *K*-detected galaxies that are located within the ALMA primary beams but $>1\farcs6$ away from an ALMA-detected SMG. The distributions of all properties are normalized to the first property ($z_{\rm phot}$) to appreciate the difference. The lower part of each panel shows the cumulative distribution and reports the K-S statistic for the corresponding properties. The photometric redshift, absolute *H*-band magnitude, and near-infrared colors appear to have the most diagnostic power to separate these two populations, although all of the properties have a significance level of the K-S statistic $<10^{-7}$, which means that the cumulative distribution function of SMGs is significantly different from that of non-SMGs. The SMGs tend to lie at higher redshift and are brighter in the rest-frame *H*-band and redder in the near-infrared colors. There are less distinct differences between the optical and ultraviolet color distributions for the SMGs and non-SMGs (in part because the reddest SMGs are not included in these plots). The final panel shows the spatial offset between the SCUBA-2 submillimeter source positions and *K*-band galaxies. This shows that we cannot simply use the spatial offset from the single-dish source position to classify SMGs and non-SMGs because of the large overlap between these two populations in terms of their spatial distributions.

flux of 3.8 ± 0.5 mJy, comparable to that detected by SCUBA-2. This indicates that the sample of "blank-ALMA" maps is dominated by real submillimeter sources.

We divide the "blank-ALMA" maps into five bins according to their SCUBA-2 flux density to further check the influence of the FPR of SCUBA-2 sources. We stack the SPIRE maps at the position of these maps separately and detect the emission in all SPIRE bands in all cases with flux densities 7–20 mJy. We also note that stacking the SPIRE images of the faintest 10% of the SCUBA-2 sources with "blank-ALMA" maps yields detections at 250 and 350 μm. This confirms that the majority of the SCUBA-2 sources that correspond to "blank-ALMA" maps are real and that our estimate of 2% false-positive sources in the parent SCUBA-2 sample is probably reasonable. The non-detection of SMGs with ALMA in these regions may be due to multiplicity (Hodge et al. 2013; Karim et al. 2013).

We will show the results of stacking the "blank-ALMA" maps at the position of machine-learning-identified SMGs in Section 4, which confirms that there are faint SMGs in these maps. Therefore, to ensure a clear separation between SMG and non-SMG samples, we do not include the *K*-band galaxies within the "blank-ALMA" maps in the "non-SMG" sample when identifying the characteristic properties of SMGs (Figure 4) or for our training set, since they may include a disproportionate number of galaxies just below our ALMA detection limit (as we show later).

3.2. Identifying the Characteristic Properties of SMGs

Having constructed clean samples of SMGs and "non-SMGs," we next compare the multiwavelength properties of these two populations to identify those properties to be used in the machine-learning analysis. We show the distributions of redshift, absolute *H*-band magnitude, and optical and near-infrared colors for ALMA SMGs and non-SMG field galaxies in Figure 4. We also present the results of Komolgorov–Smirnov (K-S) tests between the two populations for each of these observables. This figure demonstrates that photometric redshift, absolute *H*-band magnitude, and near-infrared colors are particularly effective at distinguishing SMGs from non-SMG galaxies. It is also clear from Figure 4 that those non-SMGs and SMGs detected in bluer filters show less difference in optical and ultraviolet colors, mostly as a result of the exclusion of the redder SMGs from these plots (which require a detection in at least one of the two filters used). For this reason, previous attempts to photometrically select SMG counterparts have also focused on near-infrared color selection or optical/near-infrared colors (e.g., Smail et al. 1999; Frayer et al. 2004; Yun et al. 2008; Michałowski et al. 2012; Alberts et al. 2013; Chen et al. 2016). However, as shown in Figure 4, although there are clear differences between the distributions of SMGs and non-SMGs in many properties, the overlap in any individual property is substantial. Nevertheless, as we will

show, the contamination from field galaxies can be efficiently reduced by combining optical/near-infrared colors, photometric redshifts, and absolute rest-frame H-band magnitudes.

The choice of which properties to use to most efficiently separate SMGs from non-SMGs for the machine-learning analysis has to balance two competing factors: precision and completeness. We have defined the precision in Section 2.3.1. Completeness is the number of recovered ALMA SMGs over the total number of ALMA SMGs within the overlapped region of UKIDSS and IRAC. Since including more features in the comparison is likely to yield a more precise separation, we start by using photometric redshift, absolute H-band magnitude (M_H), $(z-K)$, $(J-K)$, $(K-[3.6])$, and ([3.6]–[4.5]) (Figure 4). However, this yields a completeness of only 43% of ALMA SMGs that have all six of these features (as shown in Figure 3). Hence, to increase the completeness, we remove the $(z-K)$ color, which allows us to employ 57% of the full sample. We note that the precision of our identification is not affected by this choice, since the SMGs that are red in $(z-K)$ also tend to be red in the other three near-infrared colors. In fact, the precision of the identification increases by about 1%, which maybe be due to the enlarged sample size.

Therefore, the features that we selected for our machine-learning classification system are photometric redshift ($z_{\rm phot}$), absolute H-band magnitude (M_H), and three near-infrared colors: $(J-K)$, $(K-[3.6])$, and ([3.6]–[4.5]). We find that 69% of the ALMA-detected SMGs lying within the UKIDSS/IRAC footprint, which have K-band counterparts, have secure measurements of all of these five properties (Figure 3).

The completeness will be increased if we use fewer properties in our machine-learning analyses. However, the precision of classification decreases to just ∼50% if we only use one near-infrared color as the input feature. Therefore, we select the K-band-detected galaxies, which have secure measurements of at least two near-infrared colors to construct the training set. The selection of photometric redshift and absolute H-band magnitude does not affect the sample size because sources with detection in three near-infrared bands (and limits/detections in the other bands) all have estimated photometric redshifts in our K-selected sample (W. Hartley et al. 2018, in preparation). Removing the requirement of a secure detection at the J band or 4.5 μm modestly increases the fraction of ALMA SMGs with K-band counterparts that could be used for machine-learning analysis to 76%. In this work, we seek to develop a more complete and robust method of identifying counterparts of SMGs that are bright in several bands. This will enable us to reliably derive the physical properties of at least a subset of the SMG population. For the rest of the SMGs that are only detected in the submillimeter band or that have detected counterparts in just one or two other bands, we can learn little about their physical properties.

3.3. Radio + Machine-learning Identifications

We construct a training set that includes the ALMA SMGs and non-SMG field galaxies with the selected measurements in the UDS. We then train the machine-learning algorithms with these selected properties and build classifiers that can optimally distinguish the two different classes from the training set and hence predict the counterparts to the SMGs from the test sample.

3.3.1. The Machine-learning Method

Having selected five properties that are likely to have the diagnostic power to differentiate SMGs from the non-SMGs, we first use the SVM model to build a nonlinear classifier for optimally separating these two populations. This is implemented by using the algorithm coded in the *scikit-learn*[18] Python package (Pedregosa et al. 2011). The SVC takes a labeled training set (in this case, "SMG" versus "non-SMG") and associated set of feature vectors (e.g., observable colors) and attempts to build hyperplanes that maximize the separation between the two classes in the n-dimensional (in this case, $n=5$) feature space. Having established the hyperplane(s), new, unlabeled test data can be presented to the trained classifier to determine which class they belong to according to their relative position in this five-dimensional parameter space.

We note that the classification cannot be performed using the SVC if an object has a missing feature. This occurs if we have only a limit on the color of $(J-K)$ or ([3.6]–[4.5]) due to the lack of a secure detection at the J band or 4.5 μm. Unfortunately, there are a number of possible causes for the lack of J or 4.5 μm detection, including dust reddening, geometry, star formation history, and redshift. Therefore, we prefer not to predict these missing values through the statistical imputation algorithms (e.g., Pelckmans et al. 2005). Instead of mixing the observable properties with predicted values, we test the influence of sources with missing values using a second machine-learning model, XGBoost, which has the capacity of performing classification with missing values.

We then train the SVM classifier through a training set that includes ALMA SMGs and non-SMG K-band galaxies, which have the secure measurement of five selected properties within the ∼50 arcmin^2 area covered by our ALMA maps in the UDS. In total, 334 ALMA SMGs and 1271 non-SMGs that have secure measurements of our five selected properties are utilized to construct the training set.

We optimize the classifier parameters via k-fold cross-validation (Kohavi et al. 1995). Here we use $k=5$; i.e., we randomly divide the training set into five equally sized "folds." The classifier is trained on $k-1$ folds and validated on the remaining fold. We use the recovery rate (also called the true-positive rate (TPR), recall, or sensitivity in statistics), FPR (also referred to as the false-alarm rate or 1−specificity), and precision (also called positive predictive value) as the evaluation metrics to optimize the parameters of the SVM classifier. The recovery rate is the ratio between the number of correctly classified SMGs and the total number of ALMA SMGs in the data set. The FPR is the number of objects incorrectly classified as SMGs over the total number of non-SMGs in the data set. We defined the precision in Section 2.3.1 as the ratio between the number of correctly identified SMGs and the total number of predicted SMGs by the classifier. An optimized classifier will maximize the recovery rate and precision while simultaneously minimizing the FPR.

The SVM classifiers use a "kernel" to efficiently compute the dot product between two vectors in feature space (i.e., a similarity measure) and build a decision function that is analogous to defining a "decision" energy resulting from placing a kernel at the position of the observed properties of a source (Cristianini & Shawe-Taylor 2000). The fivefold cross-validation shows that the most efficient kernel function for separating SMGs from K-band-detected galaxies is the

[18] http://scikit-learn.org

Table 1
UDS Training Set for Machine-learning Models

Label	$z_{\rm phot}$	M_H	$(J-K)$	$(K-[3.6])$	([3.6]–[4.5])
1[a]	3.56	−24.59	2.35	0.73	0.50
1	2.50	−24.05	2.87	0.96	0.31
1	4.19	−24.34	⋯	1.25	0.27
1	3.10	−24.22	3.18	1.16	0.60
0	0.64	−21.22	1.36	−0.14	−0.33
0	0.35	−21.34	1.46	−0.48	0.17
0	2.90	−21.91	1.49	0.11	0.15
0	0.95	−23.05	1.88	0.63	−0.18
0	0.42	−18.27	1.16	−0.68	⋯

Note.
[a] SMGs are labeled as 1 and non-SMGs are labeled as 0.

(This table is available in its entirety in machine-readable form.)

polynomial kernel, which is defined as

$$k(x, x') = (\gamma(x \cdot x') + c_0)^d, \quad (2)$$

where x and x' represent feature vectors in the input space, $(x \cdot x')$ is their inner product, d denotes the degree of the polynomial kernel function, and c_0 is a constant coefficient that is an independent parameter in kernel function. The other two parameters of the SVM algorithms with a polynomial kernel are γ and C, where γ represents the adjustable kernel width parameter, which is responsible for the topology of the decision surface, and C sets the width of the margin separation of different classes of objects (e.g., Małek et al. 2013; Kurcz et al. 2016). The fivefold cross-validation shows that the default value of these parameters in the *scikit-learn* package, $C = 1.0$, $\gamma = 1/n$ features (here $n = 5$), $d = 3$, and $c_0 = 0.0$, are optimized for performing the classification in our work by an SVM classifier with a polynomial kernel.

The feature selection module in the *scikit-learn* package can also select the best features for classification based on univariate statistical tests (Pedregosa et al. 2011). The univariate score is derived by $U_{\rm score} = -\log(p)$, where p is the p-value of the corresponding univariate feature (Pedregosa et al. 2011). Among the five features we selected, the best one for separating SMGs from non-SMGs is $(J-K)$ color with a $U_{\rm score} = 891$, followed by ([3.6]–[4.5]) color with a score of 707, $(K-[3.6])$ with a score of 695, and the absolute H-band magnitude with $U_{\rm score} = 579$. The photometric redshift has a relatively lower univariate score of 324; however, as we described above, including photometric redshift and absolute H-band magnitude as the input features for machine learning does not affect the completeness of our analyses but increases the recovery rate of the SVM classifier by about 6%.

The sample we used for performing the SVM machine-learning classification are K-band-detected galaxies that have secure measurements of all five selected properties. To increase the completeness, we also include objects that lack a secure detection at the J band (i.e., have a limit on their $(J-K)$ color) or at 4.5 μm (i.e., have a limit on the ([3.6]–[4.5]) measurement). This increases the sample size of the training set from 1605 to 1832, in which 366 are ALMA SMGs and 1466 are non-SMGs. The training set we use in our analysis is given in Table 1.

As a test of the efficiency of the SVM classifier, we have also applied a second machine-learning classifier to our sample. This is XGBoost[19] (Chen & Guestrin 2016), which is a scalable machine-learning system for tree boosting. In this tree ensemble model, the input features will be first divided into different "leaves." Then, the algorithm computes the optimal weight of each "leaf" and calculates the corresponding optimal value, which will be used as a quality score of a tree structure. The structure of a tree is built by a greedy algorithm that starts from a single leaf and iteratively adds branches to the tree. Instead of enumerating all possible tree structures, XGBoost first calculates a gain of a "leaf." If the gain of the corresponding leaf is smaller than the minimum loss reduction (γ), the branch will not be added to the tree. One of the key problems in tree classifiers is how to find the best split at each node (in this case, "SMG" versus "non-SMG"). XGBoost finds the best solution among all possible splits based on the aggregated statistics according to percentiles of feature distribution. For the missing value, XGBoost classifies the instance into the optimal default direction that is learned from the data. The input properties of unlabeled test data will be divided into the same leaves as the training set, and the final prediction will be calculated by summing up the score in the corresponding "leaves" of a test object (Chen & Guestrin 2016).

Similarly, we optimize the parameters of the XGBoost tree classifier via the fivefold cross-validation. Unlike the SVM implemented in the *scikit-learn* package, which directly predicts the class label of an object, the XGBoost classifier estimates a probability of an object being an SMG. We then also use the area under the curve (AUC) of a receiver operating characteristic (ROC) curve (Fawcett 2004), as well as the assessment metrics—recovery rate, precision, and FPR—we used before to optimize the parameters of the XGBoost classifier. The ROC curves are constructed by comparing the recovery rate against the FPR, as the probability threshold is varied. Typically, an AUC higher than 0.9 indicates an excellent classifier (e.g., Lochner et al. 2016).

For boosting trees, we find that a learning rate $\eta = 1.0$ and a maximum number of iterations *num_round* $= 5$ are enough for performing a good classification (AUC > 0.9). The other two parameters for a binary classification are the minimum loss reduction (γ), which is required to make a further partition on a leaf node of the tree and the maximum depth of a tree. The fivefold cross-validation indicates that $\gamma = 1.0$ and a maximum depth of 6 are the optimized parameters for the XGBoost classifier. An object is classified as an SMG if the probability $\geqslant 0.5$.

For both machine-learning algorithms, we use a uniform weight for all objects and properties. We repeat the fivefold cross-validation 100 times, calculate the median and standard deviation of each metric, and present the values of these metrics for the optimized classifiers in Table 2.

3.3.2. Test 1: Self-test

To test the efficiency of our machine-learning method, first we carry out a "self-test," i.e., using all K-band galaxies within the ALMA primary beams to build a test set. The K-band galaxies in the 108 "blank-ALMA" maps are also included in the test sample, since it is not possible to know a priori which submillimeter sources will have "blank-ALMA" maps (i.e., contain no SMGs above a 4.3σ significance cut) when we

[19] https://github.com/dmlc/xgboost

Table 2
Summary of Machine-learning/Radio Combined Machine-learning Performances for the Different Tests

Method	Machine Learning[a]			Radio+Machine Learning		
Test/metrics	Recovery rate	Precision	FPR[b]	Recovery rate	Precision	FPR
SVM[c]	(77.2 ± 4.7)%	(82.0 ± 4.9)%	(4.7 ± 1.5)%	…	…	…
XGBoost[c]	(76.7 ± 4.9)%	(81.1 ± 4.3)%	(4.6 ± 1.2)%	…	…	…
AS2UDS self-test (SVM)	75.4%	66.7%	7.4%	85.3%	62.2%	10.2%
AS2UDS self-test (XGBoost)	73.8%	66.0%	7.2%	84.7%	62.2%	9.7%
AS2UDS "half–half" test	(75.1 ± 3.4)%	(64.1 ± 3.5)%	(8.3 ± 1.2)%	(86.3 ± 2.7)%	(60.7 ± 3.0)%	(10.1 ± 1.9)%
12% AS2UDS test	(75.6 ± 6.1)%	(67.4 ± 7.0)%	(8.3 ± 1.2)%	(86.0 ± 5.0)%	(63.0 ± 6.6)%	(10.1 ± 1.9)%
ALESS test	61.7%	70.7%	6.5%	72.3%	65.4%	9.7%
ALESS self-test	72.3%	73.9%	6.5%	78.7%	68.5%	9.2%

Notes.
[a] Machine learning refers to the SVM unless we state that we used XGBoost.
[b] FPR, which is defined as the number of objects that are incorrectly classified as SMGs over the total number of non-SMGs in the data set.
[c] The results of fivefold cross-validation for the optimized machine-learning classifier.

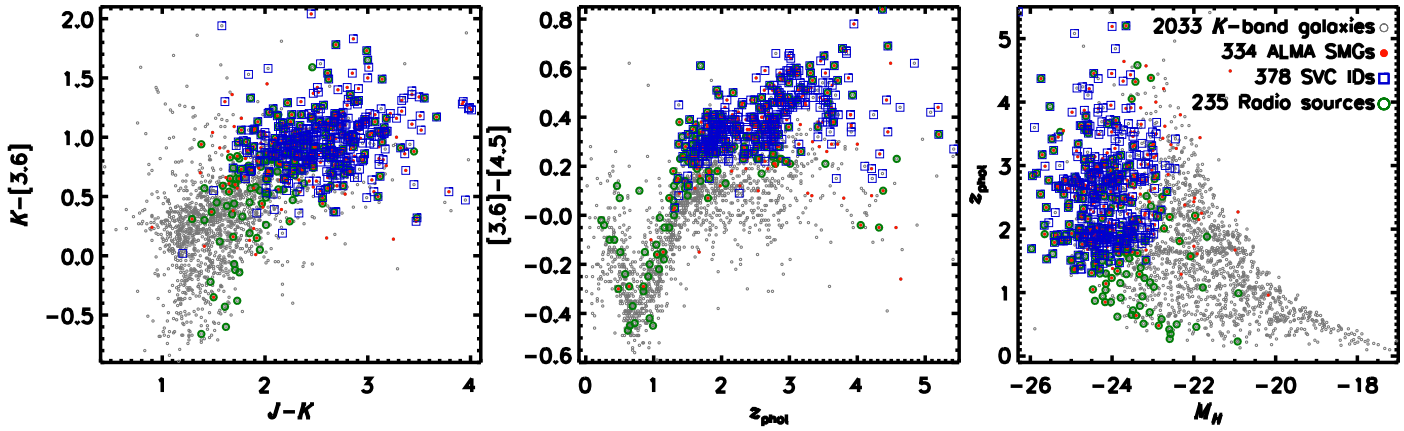

Figure 5. Results of applying the SVM-learning classifier to identify SMGs from non-SMGs to the galaxies in the UDS field, based on a training set of the full sample of ALMA-identified SMGs in AS2UDS (termed a "self-test"). We show the distributions of near-infrared colors, photometric redshift, and absolute *H*-band magnitude of 2033 *K*-band-detected galaxies lying within the ALMA maps (gray open circles). The filled points show the 334 counterparts of ALMA-detected SMGs that have secure measurements of all five observational properties. The galaxies that are classified as counterparts of SMGs by the SVC are marked by blue open squares. We also mark sources that have radio counterparts with green open circles. The SVC recovers 75% of SMGs with a precision of >67%. By including radio identifications with $p \leqslant 0.065$, the completeness of our method reaches 85% with a precision of >62%. As we have considered all *K*-band galaxies within the "blank-ALMA" maps to be non-SMGs for this test, even though our stacking results show they typically have submillimeter emission just below our detection limit, the recovery rate and precision we present in the plot should be considered as lower limits.

identify counterparts for single-dish-detected submillimeter sources. In total, 2033 *K*-band galaxies lie within the ALMA primary beams and have secure measurements of five selected properties; 363 of these are in "blank-ALMA" maps. We then utilize the training set and SVM model to identify the likely SMGs in this test sample and compare this to the actual catalog of ALMA-detected SMGs in these maps.

We present the results of the "self-test" in Figure 5. The SVC classifies 378 counterparts as "SMGs" from the 2033 *K*-band-detected galaxies within the ALMA primary beams, somewhat more than the 334 ALMA-detected SMGs in these fields. For the 334 ALMA-detected SMGs with all five features, 252/334 (75%) are recovered by the SVC model. The precision of this machine-learning method is therefore 67% (252/378). We note that this is a lower limit on the precision for the machine learning, since we consider all *K*-band galaxies in the "blank-ALMA" maps as non-SMGs. However, our stacking of far-infrared observations shows that there are faint SMGs present in the "blank-ALMA" maps, and some of machine-learning method classified "SMGs" in the "blank-ALMA" maps will be true counterparts of SMGs that are marginally too faint to be detected by ALMA (as we show later). The results of the fivefold cross-validation shown in Table 2 indicate that the precision would increase to 82% if we had excluded the "blank-ALMA" maps from the analysis.

As shown in Figure 5, those galaxies that are classified as "SMGs" by the SVM classifier but not detected by ALMA at $>4.3\sigma$ (typically $S_{870\mu m} \geqslant 0.9$ mJy) have very similar properties to the ALMA-detected SMGs; i.e., they are red in the near-infrared, at high redshift, and bright in the rest-frame *H* band. We will discuss the properties and the results of stacking the ALMA maps at the position of these galaxies in Section 4. We also note that the SMG counterparts that are not recovered by the machine-learning code tend to be those at lower redshifts, which are faint in the rest-frame *H* band or blue in their near-infrared colors.

We also highlight in Figure 5 the *K*-band galaxies that have radio counterparts with *p*-values $p \leqslant 0.065$. As we described in Section 2.3.1, we use the *p*-statistic to identify radio counterparts for single-dish-detected submillimeter sources. For the 2033 *K*-band galaxies in the UDS test sample, 235 also have $>4\sigma$ radio detections with $p \leqslant 0.065$. Among these, 167/235

(71%) are matched to ALMA-detected SMGs within 1.″6. Therefore, half of the 334 ALMA SMGs are recovered by radio identification alone. Combining the machine-learning classification with the radio identification, 285/334 (85%) of the ALMA SMGs are recovered with a precision >62%. This proves that our combined radio and machine-learning method can efficiently recover SMGs from the general population of *K*-band-selected galaxies.

To increase the completeness of the self-test sample, we also include the *K*-band-detected galaxies that lack a secure detection at the *J* band or 4.5 μm and adopt the XGBoost machine-learning module to perform the classification. The sample size is increased to 2305, with 366 being ALMA-detected SMGs. The XGBoost model identifies 409 "SMGs" from this enlarged test sample. For the ALMA SMGs, 270/366 (74%) are recovered with a precision of >66%. Combining with the radio identification, 310/366 (85%) of ALMA SMGs have been recovered with a precision of >62%.

We note that the performances of the two machine-learning modules are very similar according to the fivefold cross-validation and this self-test (Table 2). To keep the consistency with Figure 5, we show the analyses of the SVM classification in the following figures and use machine learning to refer to the SVM method, unless we explicitly state that we are using XGBoost.

3.3.3. Test 2: Independent Test

We expect the "self-test" will provide an overly optimistic indication of the success rate of our method, as it uses the same sample for both the training and testing. For that reason, we also undertake a number of independent tests, which use distinct samples for the training and testing.

First, we divide the AS2UDS sample into independent halves to test our method, which we will term a "half–half" test. We randomly assign the *K*-band galaxies in half the ALMA maps to the training set and use the galaxies within the other half of the ALMA maps as the test sample. We then utilize this training set and our combined radio and SVC machine-learning method to classify the likely counterparts of SMGs in the independent test sample. We repeat this "half–half" test 100 times to estimate the scatter in the recovery rate and precision. The median recovery rate of the combined radio and machine-learning method of these tests is 86% ± 3% with a median precision of 61% ± 3% (Figure 6). This "half–half" test confirms the success of our method when used to identify the counterparts of SMGs using a training set with similar photometric coverage and depth. This success rate is therefore expected to be representative of what will be achieved when we apply our method to identify the SMG counterparts in the S2COSMOS survey (J. M. Simpson et al. 2018, in preparation).

The next independent test we perform is to apply the trained SVM classifier to the independent sample of ALMA-identified SMGs in the ECDFS field from the ALESS survey (Hodge et al. 2013; Simpson et al. 2014). As we described in Section 2.3.3, there are 323 IRAC-selected galaxies located within the 88 ALMA primary beams. Of these galaxies, 232/323 have secure measurements of the five selected properties that are used in our SVM classification. Among these, 47/232 sources match to ALESS MAIN sample SMGs within 1.″5.

We show the results from applying the SVM classifier trained on the AS2UDS sample to the identification of the SMGs in ALESS in Figure 6. The recovery rate of the machine learning is 62% with a precision >71%. As we have included the 19 "blank-ALMA" maps in our test sample (which includes some galaxies classified as non-SMG but that are actually just below our submillimeter flux limit), we believe this precision is a lower limit. We also match these 232 IRAC-based galaxies with the radio catalog from Miller et al. (2008). Again, the radio identification can recover half of the ALMA SMGs with a precision of 75%. Hence, combining the radio identification and machine-learning classification, we recover 72% of the ALMA SMGs with a lower limit on the precision of 65%.

We note that the recovery rate for the ALESS sample is lower than that achieved in either the "self-test" or "half–half" test on the AS2UDS sample. To understand the cause of this, we also carry out a "self-test" on the ALESS sample (i.e., we use the ALESS SMGs and non-SMGs as both the training and analysis samples) and find that the recovery rate of classification increases from 72% to 79%, while the precision increases

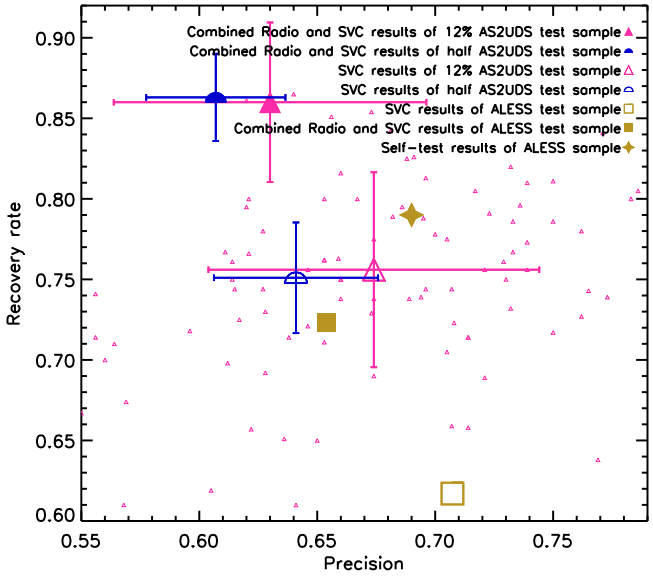

Figure 6. Results of applying the SVM-learning method to the independent samples from the AS2UDS and ALESS surveys. The open half circle shows the median precision (64% ± 3%) and recovery rate (75% ± 3%) of our machine-learning method for the "half–half" test—this involves constructing a training set from the galaxies in half of the ALMA maps in AS2UDS and testing the method on an independent sample from the other half of the ALMA maps. The filled half circle shows the results of the combined radio and machine-learning method for the "half–half" test. We show error bars estimated from the variation in the derived precision and recovery rate based on 100 bootstrap-simulated "half–half" tests. We also apply our combined radio and machine-learning method, trained on the AS2UDS sample, to the independent ALESS sample in the ECDFS field and plot this as a filled square. We recover 72% of ALESS SMGs with a precision of 65%, which can be compared to the success rate indicated by a "self-test" on the ALESS sample (filled star). To investigate the variation in recovery rate and precision as a function of sample size, we also randomly select subsamples of 88 AS2UDS ALMA maps (12% of the AS2UDS sample) matching the number of ALMA maps in ALESS. We show the results of the recovery rate (86% ± 5%) and precision (63% ± 7%) of the combined radio and machine-learning method for this "12%" test sample with a filled triangle and for just the machine learning as a large open triangle. The small open triangles represent the results of the machine-learning method for 100 individual subtests. Four of these "12%" tests have a recovery rate as low as that seen for the ALESS test sample, while the median recovery rate of these 100 tests is the same as the "self-test" of AS2UDS. These results illustrate the success rate of our combined radio and machine-learning method and the expected scatter in the recovery and precision when applied to smaller samples, including those selected from different fields from those used for the training set.

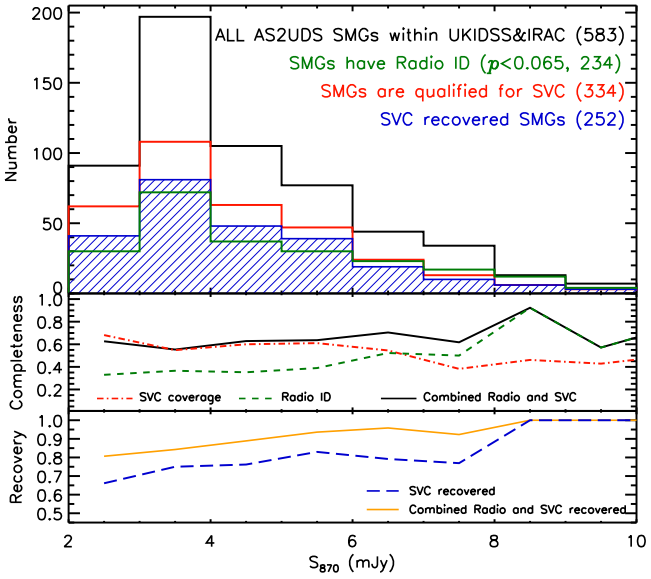

Figure 7. Recovery rate and completeness of our combined radio and SVC machine-learning methodology for identifying SMG counterparts as a function of the flux density of SCUBA-2-detected submillimeter sources ($S_{870\mu m}$). We limit our identification of counterparts to single-dish-detected submillimeter sources in the combined region of UKIDSS and IRAC, since only *K*-band galaxies within this coverage are suitable for the machine-learning method. For the SMGs that have secure measurements of the five features used to train the SVC, our method successfully recovers 85% of SMGs; of these, 77% can be recovered with just machine learning, and this fraction increases to 81% for the submillimeter sources brighter than 4.5 mJy at 870 μm. For the full sample of AS2UDS SMGs within the combined UKIDSS and IRAC coverage (not just those with the five features), 40% of SMGs have radio counterparts, and this fraction increases to 46% for brighter submillimeter sources ($S_{870\mu m} > 4.5$ mJy). Around 57% of the SMGs have the five features we use and so are qualified for our machine-learning method, and this fraction does not depend upon their submillimeter flux. Combining the radio identification and machine-learning results shows that 60% of ALMA-detected SMGs can be recovered, and this fraction increases to 71% for the brighter submillimeter sources.

to 69%. The recovery rate of ALESS SMGs is still lower than that of the "self-test" on the AS2UDS SMGs.

It may be that the lower success rate for the ALESS sample is simply due to small-number statistics: the number of ALMA maps in ALESS is only 12% (88/716) of that in AS2UDS. We can test this using AS2UDS by selecting test samples of galaxies from 88 randomly selected ALMA maps from the AS2UDS survey and determining the variation in the recovery rate and precision between these test samples. We call this the "12% test," and we repeat this test 100 times to obtain the scatter. The median recovery rate of our combined radio and machine-learning method for a sample of galaxies in 88 ALMA maps is 86% $\pm$ 5%, with a lower limit on the precision of 63% $\pm$ 7% (Figure 6). When we compare the results with the "half–half" test, we find that the smaller sample size causes larger uncertainties in the machine-learning classification. We also note that four of these 100 tests have a recovery rate as low as that of the ALESS SMGs but with a range of precisions. Therefore, it appears that a recovery rate and precision as low as those seen for the ALESS test sample are possible, simply due to the small sample size. However, we note that there are also potential astrophysical reasons for the different success rates. In particular, the ALESS SMGs are typically fainter at 870 μm, with a median flux density of $S_{870\mu m} = 2.2$ mJy, compared to the AS2UDS SMGs, $S_{870\mu m} = 3.8$ mJy. As shown in Figure 7, the recovery rate of the combined radio and machine-learning method is higher for brighter SMGs. And we note that the beam of LABOCA/APEX, which is the basis of ALESS, is larger than that of SCUBA-2/JCMT used for AS2UDS, which will reduce the precision of identification of counterparts to single-dish-detected submillimeter sources.

Thus, we argue that the decrease in the recovery rate when using the AS2UDS training set applied to the ALESS sample is probably partly caused by the relative faintness of the SMGs and larger beams of the single-dish survey in the ECDFS field. The difference between a *K*-selected training set in UDS and the IRAC-selected test sample in ECDFS may also affect the results of our method.

We also undertook these same tests using the XGBoost machine-learning classifier. The two machine-learning modules have a very similar performance on the "half–half" and "12%" tests, while the XGBoost classifier gives a marginally higher recovery rate (64%) with a relatively lower precision (65%) for the ALESS sample.

4. Results and Discussion

To determine the completeness of our method for recovering ALMA SMGs, we first summarize in Table 2 the three evaluation metrics—recovery rate, precision, and FPR—from the machine-learning classification and the combined radio and machine-learning method for SMGs in both the training set and test samples. We note that only *K*-band galaxies within the combined coverage of UKIDSS and IRAC have sufficient photometric coverage to be suitable for the machine-learning method. Hence, the completeness was defined as the ratio between the number of recovered SMGs and the total number of ALMA SMGs within this overlapped region (Section 3.2).

We also report the number of radio-detected galaxies that are located within the ALMA maps but do not have a $>4.3\sigma$ ALMA detection in Table 3. We refer to these radio sources as "radio-detected and ALMA-faint" galaxies in the following analysis, since our stacking analysis shows that they are typically just below the submillimeter detection limit of our ALMA maps. For the same reason, the *K*-band galaxies that are classified as "SMGs" by the SVC machine learning but do not have a secure ALMA detection are termed "machine-learning-classified and ALMA-faint" galaxies. To verify that, on average, there is fainter submillimeter emission from galaxies within the "blank-ALMA" maps, as suggested by their detection in the far-infrared stacking analysis in Section 3.1, we separately study the properties of radio-detected/machine-learning-classified and ALMA-faint galaxies within the "blank-ALMA" maps and the maps that contain at least one ALMA-identified SMG and list the number of each in Table 3.

4.1. Incompleteness of Our Multiwavelength IDs

We first investigate the completeness of our method for recovering ALMA SMGs in the UDS. We show the recovery rate of the combined radio and SVC machine-learning method as a function of flux density of SCUBA-2-detected submillimeter sources ($S_{870\mu m}$) in Figure 7. For the 583 AS2UDS SMGs within the overlapped region of UKIDSS and IRAC, 352/583 (60%) can be recovered by the combined radio and machine-learning method, and for the submillimeter sources brighter than 4.5 mJy at 870 μm, this fraction increases to 71%. Of the 583 ALMA-identified SMGs, 334 have secure measurements of five selected properties and therefore qualify for the SVC machine-learning

Table 3
Summary of Radio-detected or Machine-learning-classified and ALMA-faint Galaxies

Type of ALMA Maps	All ALMA Maps	Maps with ALMA ID	"Blank-ALMA" Maps
SVC-classified and ALMA-faint	**126**[a]/607[b]	**75**/512	**51**/95
Radio-detected and ALMA-faint	137/714	84/606	53/108

Notes.
[a] The number of SVC-classified/radio-detected and ALMA-faint galaxies.
[b] The number of eligible ALMA maps available.

method. The SVC successfully selected 75% of these ALMA SMGs from a sample of all *K*-band-detected galaxies within the ALMA primary beams. By including the radio identifications, the recovery rate increases to 85% (285/334).

Looking at the full SMG sample from AS2UDS within the combined UKIDSS and IRAC coverage, the radio identification alone can recover 234/583 (40%) of all ALMA-detected SMGs, and for the brighter single-dish-detected submillimeter sources ($S_{850\mu m} \geqslant 4.5$ mJy), the recovery rate increases to 49%. For the AS2UDS SMGs that do not have secure measurements of the five properties within the overlap region of the UKIDSS and IRAC observations (and to which we therefore cannot apply the SVC machine-learning method), radio identification recovered an additional 67 SMGs. For the other 231/583 (40%) ALMA SMGs that are neither qualified for the SVC machine-learning method nor have radio counterparts, it is infeasible to identify their multiwavelength counterparts. This fraction reduces to 29% for just those submillimeter sources brighter than 4.5 mJy.

The purpose of this work is to construct a training set based on a large sample of ALMA-detected SMGs and deep ancillary data in the UDS field that can then be used to identify counterparts to single-dish-detected submillimeter sources from surveys of other fields that either have not yet been observed by ALMA or cannot be observed. Therefore, as a more representative test, we also determine the completeness and recovery rate of our method when applied to independent test samples: separate halves of the AS2UDS survey sample and the ALESS survey. For the "half–half" test on AS2UDS, applying a training set constructed from half the AS2UDS maps to the galaxies in the other half of the maps, the machine-learning recovers (75% ± 3%) of SMGs with a lower limit on the precision of 64% ± 4%. When combined with the radio identifications, (86% ± 3%) of SMGs in the "half–half" AS2UDS test are recovered with a precision >61% ± 3%. For the ALESS SMGs, 47/96 (49%) of ALESS SMGs are qualified for the machine-learning method, and 29 of them were recovered. By including the radio identification, we can identify counterparts for 34/47 SMGs (72%). The radio identification recovers another 21 ALMA-detected SMGs by using the radio catalog from Biggs et al. (2011; Hodge et al. 2013; Simpson et al. 2014). In terms of the full sample of ALESS SMGs, the combination of radio and machine-learning methods yields identifications for 55/96 (57%) of the counterparts to single-dish-detected submillimeter sources in the ECDFS.

One of the limitations that causes incompleteness in our method is the fact that the SVC machine-learning method cannot deal with missing features unless they are artificially filled, as we described in Section 3.3.1. To test the effect of this limitation, we adopted the second machine-learning model, XGBoost, which has the capacity of performing classifications with the missing values. The two machine-learning algorithms have very similar performances, as we show in Table 2. The sample size for the machine-learning analysis is enlarged by including objects that lack *J*-band or 4.5 μm detection. This improved the completeness of analyses from 60% to 64%. However, for the other 46% of AS2UDS SMGs, which are only detected in the submillimeter or only have counterparts in one or two wavebands, the opportunities to learn more about their properties even if they are correctly identified are limited due to the paucity of information available on them.

4.2. Properties of ALMA-detected and ALMA-faint Radio Sources

We investigate the population of radio-detected and ALMA-faint galaxies by comparing the multiwavelength properties of ALMA-detected and ALMA-faint radio sources in the UDS. The radio imaging covers 714/716 ALMA pointings. In total, 404 radio sources fall inside the 714 ALMA primary beams. Among these, 259 match to ALMA SMGs within 1.″6 and hence are counterparts of ALMA SMGs. We define 137 radio sources as "non-SMGs," since they are located within the ALMA primary beams but >2.″6 away from ALMA SMGs or within the "blank-ALMA" maps (eight radio sources lie between 1.″6–2.″6 from the ALMA SMGs and are excluded from this analysis, as their associations are ambiguous). We show the comparison of the multiwavelength properties of these two samples of submillimeter-detected/undetected radio sources in Figure 8. We present the distribution of non-SMG radio sources lying in the ALMA maps with a detected SMG and radio sources within the "blank-ALMA" maps separately, since the far-infrared stacking analysis shows that there may be fainter submillimeter emissions from galaxies within the "blank-ALMA" maps (Section 3.1). The non-SMG radio sources within maps with an ALMA SMG tend to lie at lower redshift and are bluer in their near-infrared colors than the SMGs; i.e., they have the same properties as *K*-band-detected non-SMGs. This also confirms that our selected properties for the machine learning can efficiently separate SMGs from field galaxies. Many of the radio sources within the "blank-ALMA" maps have properties like SMGs, while some show the properties of non-SMGs.

To further investigate the radio-detected but ALMA-undetected galaxies in our field, we stack the primary-beam-corrected ALMA maps at the position of these radio sources. There are 404 radio sources located within the ~50 arcmin^2 covered by our ALMA survey. Among these, 137 are defined as non-SMGs, since they do not correspond to a >4.3σ ALMA counterpart. We separately stack the 53 radio sources that lie in the "blank-ALMA" maps and 84 non-SMG radio sources in maps with at least one ALMA-detected SMG. We show the stacked results of 53/137 radio sources within the "blank-ALMA" maps in the left panel of Figure 9. The median flux density of the stacked ALMA images is $S_{870\mu m} = (0.51 \pm 0.05)$ mJy, which is consistent with

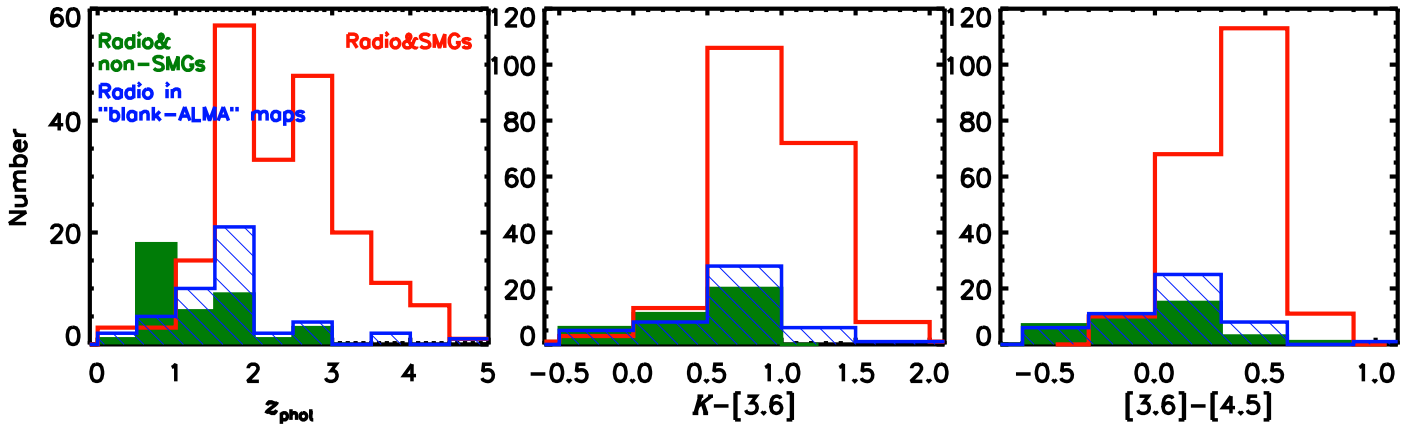

Figure 8. Distributions of photometric redshift and near-infrared colors for those radio sources that have submillimeter counterparts (SMGs) vs. those that are not individually detected by ALMA (non-SMGs). Of the 137 non-SMG radio sources lying within our ALMA coverage, 53 are located in the "blank-ALMA" maps. We show the distribution of non-SMG radio sources within the maps that contain ALMA-identified SMGs (green filled region) and the radio sources within the "blank-ALMA" maps (blue hatched region). The non-SMG radio sources lying in maps with an ALMA-detected SMG tend to lie at lower redshift and have bluer near-infrared colors. The distributions of radio sources within the "blank-ALMA" maps suggest that they may contain a mix of both higher-redshift, submillimeter sources and lower-redshift, non-SMGs, which is consistent with our stacking results, as we show later.

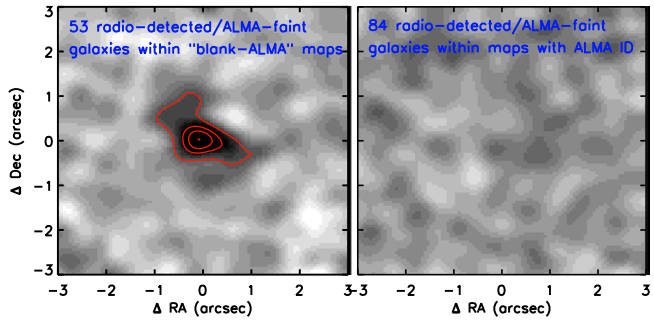

Figure 9. Results of stacking the primary-beam-corrected ALMA maps at the position of the radio sources that are individually undetected at 870 μm with ALMA. The left panel shows the average stacking results of the 53 radio sources that are in "blank-ALMA" maps (those with no detected SMGs). The median peak flux density of these galaxies is $S_{870\mu m} = (0.51 \pm 0.05)$ mJy. Contours indicate the significance of the 870 μm emission at 3σ, 6σ, 8σ, and 10σ. The right panel shows the similar stacked emission from the 84 non-SMG radio sources but now lying in the maps that contain at least one ALMA-detected SMG. The stack results confirm that these radio sources are not submillimeter sources with a 3σ limit of $S_{870\mu m} = 0.14$ mJy. However, for the radio sources within the "blank-ALMA" maps, at least, some of the radio sources are submillimeter sources, although these are just too faint (or too diffuse) to be individually detected in our ALMA observations.

the detection of significant far-infrared emission in the SPIRE stacks of these maps (Section 3.1), indicating that most of them correspond to real SCUBA-2 sources. We also stack the 84 radio sources that are individually undetected by ALMA but lie in a map with an ALMA-detected SMG. This confirms that these galaxies do not have detectable submillimeter emission, i.e., they are non-SMGs. Based on the fraction of SMGs and non-SMGs in the maps with an ALMA-identified SMG, we estimate that at least ~70% (Figure 2) of radio sources in the "blank-ALMA" maps have real submillimeter emission, although they are too faint to be detected individually in our ALMA observations.

4.3. Stacking Machine-learning-classified and ALMA-faint K-band Galaxies

As shown in Figure 5, the machine-learning-classified and ALMA-faint K-band galaxies have similar properties to the ALMA-detected SMGs: they lie at high redshift and are bright in the rest-frame H band and red in optical/near-infrared colors.

To determine whether these galaxies are submillimeter emitters that lie slightly below the detection limit of our 870 μm ALMA maps, we perform a stacking analysis at their positions in the ALMA maps. In Figure 5, we show that the machine-learning method classifies 378 "SMGs" from the 2033 K-band galaxies within the ALMA primary beams. Among these, 252 match to ALMA-identified SMGs. We show the stacked results of the other 126 SVC-classified but ALMA-undetected K-band galaxies in Figure 10. The median flux density of these 126 galaxies is $S_{870\mu m} = (0.61 \pm 0.03)$ mJy, which indicates that, on average, these sources have submillimeter emission but are too faint (or too diffuse) to be detected by our ALMA observations.

Among these 126 K-band galaxies, 51 lie in the "blank-ALMA" maps (those without an individually detected SMG). The stacked median flux density of these 51 galaxies is $S_{870\mu m} = (0.92 \pm 0.05)$ mJy. The other 75 galaxies, those in maps with an ALMA-identified SMG, have a stacked median flux density of $S_{870\mu m} = (0.42 \pm 0.04)$ mJy (Figure 10). Therefore, the machine-learning-classified "SMGs" within the "blank-ALMA" maps have submillimeter emission that is twice as bright as that of similar galaxies in the maps that already contain an individually detected SMG. This confirms our suggestion that some galaxies within the "blank-ALMA" maps have submillimeter emission just below the detection threshold of our ALMA observations. Therefore, because of the ambiguous classification of the sources within the "blank-ALMA" maps, we chose not to include them when originally constructing the training sets for the machine learning (3.3.1), as they would have blurred the distinction between the properties of the SMGs and non-SMGs. We also investigate the effect on the machine-learning training by including K-band galaxies within the "blank-ALMA" maps into the non-SMG sample of the training set. In this case, the recovery rate of SMGs based on the "self-test" decreases by about 10%. The reason for this is that in this case, we are labeling some galaxies that have the same properties as the counterparts of ALMA SMGs as "non-SMGs."

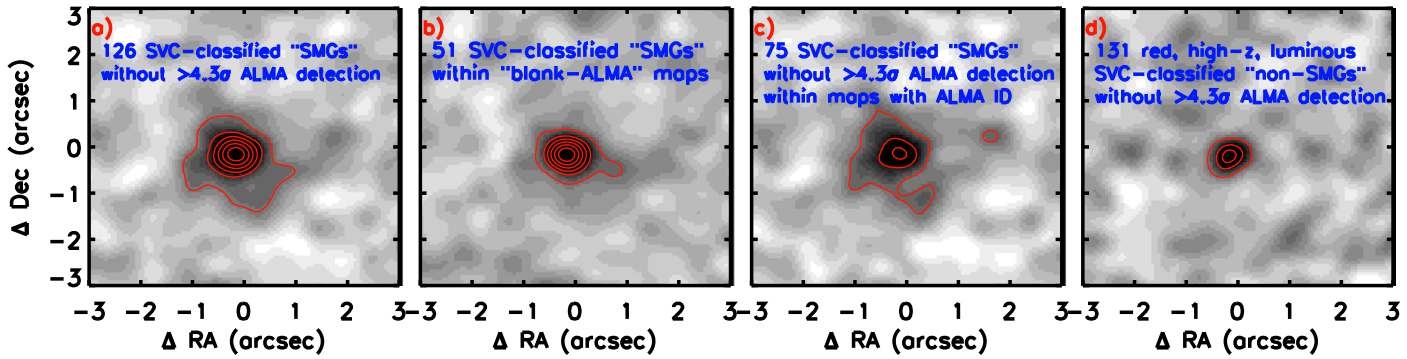

Figure 10. Results of stacking the primary-beam-corrected ALMA maps at the position of K-band galaxies that are classified as "SMG" counterparts by the machine-learning method but not individually detected above $>4.3\sigma$ in our ALMA maps. (a) Average stacking results of all 126 such galaxies. We measure a median peak flux density of $S_{870\mu m} = (0.61 \pm 0.03)$ mJy. Contours represent significance levels of 3σ, 6σ, 8σ, 10σ, 12σ, and 14σ at 870 μm. (b) Stacking results of the 51 machine-learning-classified and ALMA-faint galaxies that fall within the "blank-ALMA" maps. The median peak flux for these galaxies is $S_{870\mu m} = (0.92 \pm 0.05)$ mJy. (c) Average stacking results of the other 75 machine-learning-classified and ALMA-faint galaxies within the ALMA maps that contain a detected SMG. The median peak flux density for these sources is $S_{870\mu m} = (0.42 \pm 0.04)$ mJy. Therefore, at least on average, the stacking results confirm that these machine-learning method classified "SMGs" have detectable submillimeter emission at the ∼0.5–1 mJy level. Equally interestingly, the median flux density of the machine-learning "SMGs" within the "blank-ALMA" maps is twice that of similar "SMGs" in non-blank-ALMA maps, which confirms that the SCUBA-2 detections in these regions are likely to be real (even if no individual galaxy was detectable with ALMA). (d) Stacking results of 131 redder ($(J - K) > 2.0$, $(K - [3.6]) > 0.5$, $([3.6]–[4.5]) > 0.05$), high-redshift ($z_{\rm phot} > 1.5$), brighter ($M_H < -22.0$ mag) K-band galaxies that are classified as "non-SMGs" by machine learning. The stacking result shows that, on average, there is also fainter submillimeter emission from these galaxies. However, the peak flux density of this stacked map is $S_{870\mu m} = (0.30 \pm 0.03)$ mJy, which is half of the peak flux density of the stacked maps at the position of 126 machine-learning classified "SMGs" (which also do not have a $>4.3\sigma$ ALMA detection). This confirms that machine learning tends to pick out the brighter SMGs.

We also compare the machine-learning results to those of simple cuts on the near-infrared colors, photometric redshifts, and absolute H-band magnitudes to select the redder ($((J - K) > 2.0$, $(K - [3.6]) > 0.5$, $([3.6]–[4.5]) > 0.05)$), higher-redshift ($z_{\rm phot} > 1.5$), and brighter ($M_H < -22.0$ mag) galaxies as probable counterparts of SMGs. There are 483 K-band galaxies in the AS2UDS test sample that meet the above criteria. Among these, 251 are ALMA-detected SMGs. Therefore, the recovery rate of this simple method is similar to that of the machine-learning method (which recovers 252 ALMA-detected SMGs). However, the precision of this simple method is just 52% (251/483), 15% lower than that of the machine-learning method. Thus, while simple cuts on a small number of observables can be used to identify probable SMG counterparts, the contamination from non-submillimeter-bright galaxies is significantly worse than that achieved by the machine-learning method. In fact, from the 483 redder, higher-redshift, and brighter galaxies selected by these simple parametric cuts, 131 are classified as "non-SMGs" by the machine-learning method. Stacking the primary-beam-corrected ALMA maps at the position of these 131 K-band galaxies, we find that the peak flux density of the stacked map is $S_{870\mu m} = (0.30 \pm 0.03)$ mJy. This is half of the peak flux density of the stacked maps at the positions of the 126 machine-learning-classified "SMGs" (which also do not have a $>4.3\sigma$ ALMA detection; Figure 10). This confirms that the machine-learning approach is more effective than simple cuts in terms of identifying the counterparts of brighter submillimeter sources.

5. Conclusions

From ALMA follow-up observations of the 716 SCUBA-2-detected submillimeter sources in the S2CLS UKIDSS-UDS field (Stach et al. 2018a; S. M. Stach et al. 2018b, in preparation), we exploit a sample of 695 SMGs ($>4.3\sigma$) within 608 ALMA maps. We label the other 108 ALMA maps that do not contain a $>4.3\sigma$ ALMA SMG as "blank-ALMA" maps. Utilizing our high-resolution ALMA data, we first identify radio, optical, and near-infrared counterparts to the ALMA SMGs. We define as a "non-SMG" any radio/K-band galaxies that are located within the primary beams of our ALMA maps but are $>2\farcs6$ (radio non-SMGs) or $>1\farcs6$ (K-band non-SMGs) away from an ALMA-detected SMG. Based on the samples of ALMA SMGs and non-SMGs, we develop a combined radio and machine-learning method using SVC to identify multi-wavelength counterparts to the single-dish-detected submillimeter sources. The main conclusions from our work are as follows.

1. We identify radio counterparts to the ALMA SMGs in the UDS. In total, there are 404 radio sources within the primary-beam coverage of our ALMA maps. Out of 695 ALMA SMGs, 268 match to 259 radio sources within $1\farcs6$. We adopt a p-value cut to identify radio counterparts to single-dish submillimeter sources. We consider 363 of the 404 radio sources with $p < 0.065$ as counterparts to SMGs. Among them, 254 are matched to 263 ALMA SMGs within $1\farcs6$. The radio identification step can recover 37% of SMGs from the single-dish survey in the UDS with a precision >70%.
2. We identify optical/near-infrared counterparts by matching the ALMA SMGs to a deep K-band-detected photometric catalog. Within the overlap region of UKIDSS and IRAC coverage, 483 K-band galaxies match to ALMA SMGs within $0\farcs6$. Therefore, ∼83% (483/583) of the ALMA SMGs in this region have K-band counterparts. We find that the photometric redshift, absolute rest-frame H-band magnitude, and near-infrared colors ($(J - K)$, $(K - [3.6])$, and $([3.6]–[4.5])$) of these SMGs appear to provide the most diagnostic power to differentiate SMGs from non-SMGs. We construct a training set that includes ALMA SMGs and non-SMG K-band galaxies with secure measurements of these five selected properties. We do not include those K-band galaxies within the "blank-ALMA" maps in the non-SMG sample used in the training set, since our stacking results indicate that these sources are faint submillimeter emitters.

3. We train the SVC machine-learning classifier using a training set of SMGs and non-SMGs and then classify the sources in a test sample. We perform a "self-test" of our machine-learning method by classifying all of the 2033 *K*-band galaxies that have secure measurements of the five selected properties and are within the ALMA primary beams in the UDS. Among these, 334 are AS2UDS ALMA-detected SMGs. The machine learning classifies 378 *K*-band galaxies as the counterparts of SMGs with a recovery rate of 75% and a precision of 67%. Our stacking results show that there is faint submillimeter emission just below our ALMA detection threshold from the galaxies that are classified as "SMGs" (but are not ALMA-detected) by the SVC machine-learning method. Therefore, both the recovery rate and precision of the machine-learning method should be considered as the lower limits. Combined with the radio identification, our method can recover >85% SMGs that have secure measurements of five selected features with a precision of >62%.
4. To test our method, we use a training set constructed from the galaxies in a randomly selected half of our AS2UDS ALMA maps to an independent test sample from the other half of the ALMA maps. We estimate a recovery rate of 86% ± 3% and a precision of >61% ± 3% from this "half–half" test, confirming the robustness of our method of identifying counterparts for single-dish-detected submillimeter sources when using a training set from the same field. We also apply our method from a *K*-detected training set in the UDS field to the IRAC-detected galaxies in the ECDFS field to predict counterparts to LABOCA-detected submillimeter sources. We use the ALMA-detected SMGs in this field from the ALESS survey to check the recovery rate and precision of our method. The combined radio and machine-learning method recovers 72% of ALMA SMGs with a lower limit on the precision of 65%. We show that the decrease of recovery rate is likely to be partly the relative faintness of the ALESS SMGs and larger beam of APEX/LABOCA, compared to those in the AS2UDS training set. The difference between the *K*-band-detected training set in the UDS field and the IRAC-detected test sample in the ECDFS field may also affect the precision of our method. We also show that the smaller sample size of ALESS causes increased uncertainties in the classifications.
5. The main limitation of our method is that we miss the SMGs that do not qualify for machine learning or have a radio counterpart. We estimate the fraction of missed sources by checking the recovery rate of ALMA SMGs. In the overlapped region of ALMA, UKIDSS, IRAC, and VLA, 60% of ALMA SMGs in the UDS field are recovered by our combined radio and machine-learning method. This fraction increases to 71% for SMGs brighter than $S_{850\mu m} \geqslant 4.5$ mJy. The completeness of recovered SMGs increases to 64% if we adopt a second machine-learning module, XGBoost. This machine-learning algorithm has a very similar performance to SVM in classifying SMGs from *K*-band-detected field galaxies but can deal with missing features, so it can employ a slightly large test sample by including objects with only limits on their *J*-band or 4.5 μm fluxes.
6. By stacking the emission in ALMA maps at the position of the machine-learning classified but individually ALMA-undetected *K*-band galaxies, we show that, on average, there is faint submillimeter emission from these galaxies. Moreover, a stack of the far-infrared *Herschel* SPIRE maps at the position of the "blank-ALMA" maps and a stack of "blank-ALMA" maps at the position of machine-learning classified "SMGs" within these maps demonstrate that the majority of SCUBA-2 sources are real, although the SMGs responsible for these sources are too faint and/or diffuse to be detected by ALMA.

In summary, the combined radio and machine-learning technique developed in this work can be used to construct large samples of likely SMG counterparts from wide-field single-dish submillimeter surveys that currently lack interferometric submillimeter follow-up, such as the remaining fields in the S2CLS (Geach et al. 2017) or S2COSMOS (J. M. Simpson et al. 2018, in preparation). These statistically large samples will enable us to investigate science questions related to the formation of SMGs and their evolutionary connections with other populations, such as high-redshift QSOs, compact red galaxies at $z \sim 1$–3, and, ultimately, massive galaxies at $z \sim 0$.

We publish the training set of SMG and non-SMG sources from the AS2UDS survey as a machine-readable catalog with this paper to allow others to apply the machine-learning method we adopted in this work to other fields.

F.X.A. acknowledges support from the China Scholarship Council for studying two years at Durham University. All Durham coauthors acknowledge STFC support through grant ST/P000541/1. I.R.S., B.G., and E.A.C. acknowledge the ERC Advanced Grant DUSTYGAL (321334). I.R.S. also acknowledges a Royal Society/Wolfson Merit Award. F.X.A. also acknowledges support from the National Key Research and Development Program of China (No. 2017YFA0402703) and NSFC grant (11773076). J.L.W. acknowledges the support of an Ernest Rutherford Fellowship. J.E.G. acknowledges the Royal Society. F.X.A acknowledges Ryley Hill for helpful discussions about the machine-learning algorithms. We thank the staff at UKIRT for their efforts in ensuring the success of the UDS project. The James Clerk Maxwell Telescope has historically been operated by the Joint Astronomy Centre on behalf of the Science and Technology Facilities Council of the United Kingdom, the National Research Council of Canada, and the Netherlands Organisation for Scientific Research. Additional funds for the construction of SCUBA-2 were provided by the Canada Foundation for Innovation. This paper makes use of the following ALMA data: ADS/JAO.ALMA#2012.1.00090.S, 2015.1.01528.S, and 2016.1.00434.S. ALMA is a partnership of the ESO (representing its member states), NSF (USA), and NINS (Japan), together with the NRC (Canada), NSC and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by the ESO, AUI/NRAO, and NAOJ.

ORCID iDs

Fang Xia An https://orcid.org/0000-0001-7943-0166
S. M. Stach https://orcid.org/0000-0003-1122-6948
Ian Smail https://orcid.org/0000-0003-3037-257X
A. M. Swinbank https://orcid.org/0000-0003-1192-5837
O. Almaini https://orcid.org/0000-0001-9328-3991

R. J. Ivison https://orcid.org/0000-0001-5118-1313
J. L. Wardlow https://orcid.org/0000-0003-2376-8971
E. A. Cooke https://orcid.org/0000-0003-3843-8393
Chian-Chou Chen https://orcid.org/0000-0002-3805-0789
J. M. Simpson https://orcid.org/0000-0002-8521-1995
D. Scott https://orcid.org/0000-0002-6878-9840
D. Farrah https://orcid.org/0000-0003-1748-2010
P. van der Werf https://orcid.org/0000-0001-5434-5942
A. W. Blain https://orcid.org/0000-0001-7489-5167
C. Conselice https://orcid.org/0000-0003-1949-7638
K. E. K. Coppin https://orcid.org/0000-0002-0729-2988

References

Alberts, S., Wilson, G. W., Lu, Y., et al. 2013, MNRAS, 431, 194
Aravena, M., Decarli, R., Walter, F., et al. 2016, ApJ, 833, 68
Barger, A. J., Cowie, L. L., Sanders, D. B., et al. 1998, Natur, 394, 248
Barger, A. J., Cowie, L. L., Smail, I., et al. 1999, AJ, 117, 2656
Barger, A. J., Wang, W.-H., Cowie, L. L., et al. 2012, ApJ, 761, 89
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Biggs, A. D., Ivison, R. J., Ibar, E., et al. 2011, MNRAS, 413, 2314
Blain, A. W., & Longair, M. S. 1993, MNRAS, 264, 509
Bloom, J. S., Richards, J. W., Nugent, P. E., et al. 2012, PASP, 124, 1175
Bothwell, M. S., Smail, I., Chapman, S. C., et al. 2013, MNRAS, 429, 3047
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Casey, C. M., Narayanan, D., & Cooray, A. 2014, PhR, 541, 45
Chapin, E. L., Chapman, S. C., Coppin, K. E., et al. 2011, MNRAS, 411, 505
Chapman, S. C., Blain, A. W., Smail, I., & Ivison, R. J. 2005, ApJ, 622, 772
Chen, C.-C., Smail, I., Ivison, R. J., et al. 2016, ApJ, 820, 82
Chen, T., & Guestrin, C. 2016, arXiv:1603.02754
Cimatti, A., Cassata, P., Pozzetti, L., et al. 2008, A&A, 482, 21
Coppin, K., Chapin, E. L., Mortier, A. M. J., et al. 2006, MNRAS, 372, 1621
Coppin, K. E. K., Swinbank, A. M., Neri, R., et al. 2008, MNRAS, 389, 45
Cowie, L. L., Barger, A. J., Hsu, L.-Y., et al. 2017, ApJ, 837, 139
Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods (Cambridge: Cambridge Univ. Press)
Danielson, A. L. R., Swinbank, A. M., Smail, I., et al. 2017, ApJ, 840, 78
Downes, A. J. B., Peacock, J. A., Savage, A., & Carrie, D. R. 1986, MNRAS, 218, 31
du Buisson, L., Sivanandam, N., Bassett, B. A., & Smith, M. 2015, MNRAS, 454, 2026
Dunlop, J. S., McLure, R. J., Biggs, A. D., et al. 2017, MNRAS, 466, 861
Dunlop, J. S., Peacock, J. A., Savage, A., et al. 1989, MNRAS, 238, 1171
Fawcett, T. 2004, Machine Learning, 31
Frayer, D. T., Reddy, N. A., Armus, L., et al. 2004, AJ, 127, 728
Friedman, J. 2001, AnSta, 29, 1189
Furusawa, H., Kosugi, G., Akiyama, M., et al. 2008, ApJS, 176, 1
Garclía-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. 2010, Neural Comput & Applic, 19, 263
Geach, J. E., Chapin, E. L., Coppin, K. E. K., et al. 2013, MNRAS, 432, 53
Geach, J. E., Dunlop, J. S., Halpern, M., et al. 2017, MNRAS, 465, 1789
Hainline, L. J., Blain, A. W., Smail, I., et al. 2009, ApJ, 699, 1610
Hartley, W. G., Almaini, O., Mortlock, A., et al. 2013, MNRAS, 431, 3045
Hill, R., Chapman, S. C., Scott, D., et al. 2018, MNRAS, 477, 2042
Hodge, J. A., Karim, A., Smail, I., et al. 2013, ApJ, 768, 91
Holland, W. S., Bintley, D., Chapin, E. L., et al. 2013, MNRAS, 430, 2513
Hughes, D. H., Serjeant, S., Dunlop, J., et al. 1998, Natur, 394, 241
Ikarashi, S., Kohno, K., Aguirre, J. E., et al. 2011, MNRAS, 415, 3081
Ivison, R. J., Greve, T. R., Dunlop, J. S., et al. 2007, MNRAS, 380, 199
Ivison, R. J., Greve, T. R., Smail, I., et al. 2002, MNRAS, 337, 1
Ivison, R. J., Magnelli, B., Ibar, E., et al. 2010, A&A, 518, L31
Ivison, R. J., Smail, I., Le Borgne, J.-F., et al. 1998, MNRAS, 298, 583
Jarvis, M. J., Bonfield, D. G., Bruce, V. A., et al. 2013, MNRAS, 428, 1281
Karim, A., Swinbank, A. M., Hodge, J. A., et al. 2013, MNRAS, 432, 2
Kohavi, R., et al. 1995, in Proc. Int. Joint Conf. on Artificial Intelligence 14 Vol. 2 (San Francisco, CA: Morgan Kaufmann Publishers), 1137
Kurcz, A., Bilicki, M., Solarz, A., et al. 2016, A&A, 592, A25
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Lilly, S. J., Eales, S. A., Gear, W. K. P., et al. 1999, ApJ, 518, 641
Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31
Małek, K., Solarz, A., Pollo, A., et al. 2013, A&A, 557, A16
Michałowski, M. J., Dunlop, J. S., Ivison, R. J., et al. 2012, MNRAS, 426, 1845
Miller, N. A., Fomalont, E. B., Kellermann, K. I., et al. 2008, ApJS, 179, 114
Mortlock, A., Conselice, C. J., Hartley, W. G., et al. 2013, MNRAS, 433, 1185
Oke, J. B. 1974, ApJS, 27, 21
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pelckmans, K., Brabanter, J. D., Suykens, J. A. K., & Moor, B. D. 2005, NN, 18, 684
Pope, A., Scott, D., Dickinson, M., et al. 2006, MNRAS, 370, 1185
Scott, K. S., Wilson, G. W., Aretxaga, I., et al. 2012, MNRAS, 423, 575
Scott, S. E., Fox, M. J., Dunlop, J. S., et al. 2002, MNRAS, 331, 817
Simpson, C., Westoby, P., Arumugam, V., et al. 2013, MNRAS, 433, 2647
Simpson, J. M., Smail, I., Swinbank, A. M., et al. 2015a, ApJ, 807, 128
Simpson, J. M., Smail, I., Swinbank, A. M., et al. 2015b, ApJ, 799, 81
Simpson, J. M., Smail, I., Swinbank, A. M., et al. 2017, ApJ, 839, 58
Simpson, J. M., Swinbank, A. M., Smail, I., et al. 2014, ApJ, 788, 125
Smail, I., Ivison, R. J., & Blain, A. W. 1997, ApJL, 490, L5
Smail, I., Ivison, R. J., Blain, A. W., & Kneib, J.-P. 2002, MNRAS, 331, 495
Smail, I., Ivison, R. J., Kneib, J. -P., et al. 1999, MNRAS, 308, 1061
Smolčić, V., Aravena, M., Navarrete, F., et al. 2012, A&A, 548, A4
Solarz, A., Pollo, A., Takeuchi, T. T., et al. 2012, A&A, 541, A50
Stach, S. M., Smail, I., Swinbank, A. M., et al. 2018a, ApJ, 860, 161
Swinbank, A. M., Chapman, S. C., Smail, I., et al. 2006, MNRAS, 371, 465
Swinbank, A. M., Dye, S., Nightingale, J. W., et al. 2015, ApJL, 806, L17
Swinbank, A. M., Simpson, J. M., Smail, I., et al. 2014, MNRAS, 438, 1267
Thomson, A. P., Ivison, R. J., Simpson, J. M., et al. 2014, MNRAS, 442, 577
Toft, S., Smolčić, V., Magnelli, B., et al. 2014, ApJ, 782, 68
Vapnik, V. N. 1995, The Nature of Statistical Learning Theory (Berlin: Springer)
Walter, F., Decarli, R., Aravena, M., et al. 2016, ApJ, 833, 67
Wang, W.-H., Chen, H.-W., & Huang, K.-Y. 2012, ApJL, 761, L32
Wang, W.-H., Cowie, L. L., Barger, A. J., & Williams, J. P. 2011, ApJL, 726, L18
Wang, W.-H., Lin, W.-C., Lim, C.-F., et al. 2017, ApJ, 850, 37
Wardlow, J. L., Cooray, A., Osage, W., et al. 2017, ApJ, 837, 12
Wardlow, J. L., Smail, I., Coppin, K. E. K., et al. 2011, MNRAS, 415, 1479
Weiß, A., Kovács, A., Coppin, K., et al. 2009, ApJ, 707, 1201
Whitaker, K. E., Kriek, M., van Dokkum, P. G., et al. 2012, ApJ, 745, 179
Yun, M. S., Aretxaga, I., Ashby, M. L. N., et al. 2008, MNRAS, 389, 333
Yun, M. S., Scott, K. S., Guo, Y., et al. 2012, MNRAS, 420, 957